Finding $H\alpha$ emission line source candidates in large photometric surveys

A thesis submitted in partial fulfilment of the requirements for the Degree

of Master of Science in Physics

at the University of Canterbury

by Claudio Schill

University of Canterbury

2019

School of Physical and Chemical Sciences

Supervisors: Assistant Professor Simone Scaringi, Associate Professor Karen Pollard

Abstract

Many point sources that are of interest, particularly for stellar evolution, emit in-excess at the $H\alpha$ emission line of the hydrogen Balmer series. However, finding $H\alpha$ emitters spectroscopically is very expensive, in terms of both time and cost, and therefore unfeasible on a scale similar to photometric surveys. Therefore, large photometric surveys with a narrow band $H\alpha$ filter can be used to find new potential $H\alpha$ emission line objects in an efficient manner, as done by Witham et al. (2008) [66] using the INT/WFC Photometric $H\alpha$ Survey of the Northern Galactic Plane (IPHAS). Combining such a survey with a large photometric survey that includes parallax measurements and therefore distance and absolute magnitude, such as Gaia, allows selection of emitters that would otherwise be missed due to reddening. In this study density-based clustering was applied to the GAIA/IPHAS value added catalogue of Scaringi et al. (2018) [51], to identify potential emitter sources. These were then further scored against their local neighbourhood in the CMD diagram, and sources exceeding the specified thresholds were selected as $H\alpha$ emitters. The selected emitters were validated manually, against LAMOST, SIMBAD and the Witham et al. (2008) [66] $H\alpha$ emitter catalogue, which showed overall good accuracy, with a low number of incorrectly selected sources. Many of the selected $H\alpha$ emitters are not in SIMBAD or Witham et al.'s (2008) [66] catalogue, indicating that these are most likely newly found $H\alpha$ emitters. This catalogue provides a good start point for spectroscopical follow-up observations, and combined with Gaia distances and proper motions can be used to identify possible new young open clusters and regions of $H\alpha$ emitter over-densities.

Acknowledgements

I firstly would like to thank my supervisors Assistant Professor Simone Scaringi at Texas Tech University and Associate Professor Karen Pollard at University of Canterbury for supporting me and providing valuable feedback for this work. I would especially like to thank Assistant Professor Scaringi for coming up with the idea for this project and getting me started, in addition to answering many of my queries. I would also like to thank my girlfriend for her unwavering support and encouragement during this year of research and writing.

This paper makes use of data obtained as part of the INT Photometric $H\alpha$ Survey of the Northern Galactic Plane (IPHAS, www.iphas.org) carried out at the Isaac Newton Telescope (INT). The INT is operated on the island of La Palma by the Isaac Newton Group in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofisica de Canarias. All IPHAS data are processed by the Cambridge Astronomical Survey Unit, at the Institute of Astronomy in Cambridge. The bandmerged DR2 catalogue was assembled at the Centre for Astrophysics Research, University of Hertfordshire, supported by STFC grant STJ0013331. This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France. The original description of the VizieR service was published in A&AS 143, 23. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

Declaration

I hereby declare that except where specific reference is made to the work of others, the content of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification. This dissertation is my own work and contains nothing which is the outcome of work done collaboration with others, except as specified in the text and Acknowledgements.

Claudio Schill

February, 2019

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Stellar evolution covers how stars form and change and is an area of active research in astronomy. Much is known about how stars form and evolve, however there are still significant gaps in our knowledge. Identifying more systems and stars is an important step to further improve and validate our current models. Young stellar objects such as T Tauri and Herbig Ae/Be stars are relevant to further improve our understanding of star formation regions and what triggers it. Objects such as accreting compact binaries allow further study of what happens to stellar remnants. Finding more of these type of objects, allows further study of these objects and stellar evolution, in addition to gaining a better understanding of their populations and distributions. Many of these objects are not easy to find, but a large number of these produce excess emission at the $H\alpha$ hydrogen emission line; some of these $H\alpha$ emitting sources are discussed in section 1.2. $H\alpha$ is transition of the hydrogen atom, producing a spectral line at a wavelength of 656 nm; the transition corresponds to an electron dropping from the third to the second energy level. The spectral line is one of the six named spectral line emissions designated as Balmer series. The main relevant sources of $H\alpha$ emission are from accretion disks, covered in more detail in section 1.1.1, and ionised hydrogen, section 1.1.2. In addition to some of the point sources listed above, $H\alpha$ emission also occurs in HII regions, which are often associated with active star formation regions as the ionisation source often is a O or B type star. These type of stars only live for a short amount of time with respect to later spectral type stars, and hence were most likely born within the cloud.

1.1 What Causes $H\alpha$ Emission

$H\alpha$ emission is primarily caused by two phenomena; accretion disks and ionisation of hydrogen.

1.1.1 Accretion Disks

Accretion disks form when matter falls towards a massive central object, such as a star or black hole. As the in-falling matter has angular momentum it does not fall directly onto/into the central object but instead forms an accretion disk around the object due to conservation of angular momentum. Particles move in an approximate circular orbit around the massive central object, while slowly losing kinetic energy and angular momentum due to viscous interaction with particles of adjacent radii. Viscosity is an internal friction that converts the kinetic energy of particles into random thermal motion, causing the gas to heat up during its descent towards the central object. [29, p. 101] [12, p. 661] This loss in energy results in the particles slowly drifting to smaller and smaller radii until they reach the surface or get channeled to the surface by the magnetic field of the central object. Neglecting the gravitational self-binding energy of the disk and assuming the disk radiates as a black body, the temperature of the disk at different radii, r from the central object, can be found to be:

$$T(r) = \left(\frac{GM\dot{M}}{8\pi\sigma}\right)^{1/4} r^{-3/4}, \tag{1.1}$$

where G is the gravitational constant, M the mass of the central massive object and $\dot{M}$ the accretion rate of matter onto the disk and σ is the Stefan-Boltzman constant. The derivation of this equation can be found in Maoz (2016) [29, pp. 101–102]. In a steady-state disk, $\dot{M}$ is not a function of r and is equal to the amount of matter reaching the surface of the central object. Therefore $T \propto r^{-3/4}$, which shows that the gas of the disk gets hotter as it circles inwards and the inner regions are the hottest, producing the largest proportion of luminosity of the disk.

For an interacting binary system, with a white dwarf accreting matter from a main sequence star through the first Lagrange point L_1, known as a cataclysmic variable, the luminosity of the disk is dominated by the inner radius. A typical white dwarf of mass 1 $M_\odot$, radius 10^4 km and accretion rate 10^{-9} $M_\odot$ yr^{-1} produces a temperature of $5 \times 10^4 K$ at an inner radius of 10^9 cm using equation 1.1. This means that the blackbody spectrum peak occurs in the UV. For young stellar objects the thermal spectrum peaks in the infrared and for black holes in the X-ray part of the spectrum. For more details on accretion disks see Maoz (2016) [29, pp. 99–108] and Carroll & Ostlie (2007) [12, pp. 661–668].

Both young stellar objects and cataclysmic variable stars with accretion disks have been observed to have strong emission from the Balmer series, which includes $H\alpha$. If the central star has a weak magnetic field and the accretion disk reaches the surface, $H\alpha$ emission may be caused by the optically thin (i.e. large radius) part or the irradiated parts of the accretion disk by the central star and companion. [65, 64]

1.1.2 Ionised Hydrogen

Ionised hydrogen, referred to as HII in astronomy, occurs when a photon of sufficient energy (> 13.6 eV, i.e. in the UV and X-ray spectrum of light) collides with a neutral hydrogen atom (HI) and the electron is lost. In the recombination process, an electron recombines with an ionised hydrogen to either the ground state or an excited state. If recombination occurs to an excited state the electron will generally decay to the ground state by various intermediate decay steps and each of these will emit an emission line corresponding to the change in energy of the states. The probability of the different transitions in energy states can be calculated using quantum mechanics; with the most dominant transition being the $H\alpha$ transition from energy level $n = 3$ to $n = 2$. [12, pp. 431–432]

Star formation regions are often associated with HII regions, as these form around young massive (O and B type) stars emitting at high enough energy to ionise the hydrogen in its remaining molecular cloud. The ionisation occurs in an approximate sphere, defined by the Strömgren radius, around the excitation source, in which all photons of energy > 13.6 eV ionise hydrogen. In an equilibrium state, the volume of the ionisation region is constant, as the rate of ionisation and recombination must be equal each other. Only photons of energies greater than 13.6 eV will ionise hydrogen, for all other photons the ionised gas is largely transparent. The energy of the HII region, received from the central star, is mainly radiated away at the discrete transition wavelengths of hydrogen. A recombining hydrogen atom will recombine either into the ground state or into an excited state with the excess energy ($E_k - E_n$) emitted as a photon. As discussed above, the electron will then decay to the ground state either directly or via intermediate energy levels, with the photons from intermediate transitions, such as Balmer or Paschen series, escaping the gas. The photons emitted from the transition to the ground state, called Lyman photons, will be re-absorbed by one of the neutral hydrogens and then the process will repeat; resulting in the initial excitation energy from recombination being degraded into many lower energy transitions. The Lyman photons will escape the HII region via a random walk of absorption and reemission. For more details see Maoz (2016) [29, pp. 122–132] and Caroll & Ostlie (2007) [12, pp. 431–432].

1.2 Sources of $H\alpha$

There are many objects that emit excess $H\alpha$ emission, this section aims to provide a brief overview of the main types of sources.

1.2.1 Accreting Compact Binaries

This is a large group of stellar remnants that accrete matter from a companion star, generally via an accretion disk that forms due to the angular momentum of the in-falling matter.

Cataclysmic Variables

Cataclysmic variables are interacting binary systems with a white dwarf as the primary component accreting matter from a secondary star, usually a G or later type main-sequence star. In these systems the secondary star fills its Roche Lobe and mass is transferred to the white dwarf. Depending on the strength of the magnetic field, the matter is either transferred completely via an accretion disk; a partial accretion disc on the outside and then channelled by the magnetic field; or if the field is very strong, no accretion disk is formed and the matter is channelled directly to the magnetic poles of the white dwarf. Cataclysmic variables go through long periods of low brightness, called quiescent, followed by short periods of large outbursts in brightness, with the increase in brightness depending on the type of cataclysmic variable. The main types are:

- **Dwarf Novae**: The outburst of a dwarf nova is due to a brightening of the accretion disk surrounding the white dwarf, increasing in brightness between 2 and 6 magnitudes. As brightened visible wavelengths precedes the UV, it is thought that the outburst starts in the cooler (i.e. blackbody, larger wavelength), outer parts and proceeds to the inner regions of the accretion disk, and is caused by a sudden increase in mass flowing through the accretion disk. Outbursts usually last between 5-20 days followed by a quiescent period of 30-300 days. The source of the increase in the mass transfer is thought to either be an instability in the transfer rate from the secondary star to the primary, or an instability in the accretion disk [12, pp. 675–680].

- **Classical Novae** are characterised by higher accretion rates than dwarf novae, outbursts take a few decades to return to quiescent brightness levels (on average $M_v = 4.5$) and during outbursts the average brightness increases by ∼10-12 magnitudes. The accepted theory for the outbursts of classical novae is that the white dwarf accretes matter at a rate of 10^{-8} to 10^{-9} $M_\odot$ yr^{-1}, with the gas accumulating on the white dwarfs surface. When

enough of the hydrogen-rich gas has accumulated on the white dwarf (10^{-4} to 10^{-5} $M_\odot$), a runaway thermonuclear reaction occurs, during which luminosities exceed the Eddington limit, resulting in radiation pressure expelling some of the accreted matter into space. This is followed by a period of a few months to a year of hydrostatic burning, at the end of which the last of the accreted matter is ejected and the white dwarf starts cooling and slowly returns to quiescent levels. [12, pp. 680–686]

- **Type Ia Supernovae** reach an average maximum visible brightness of -19.3 magnitudes and have a well defined rate of decline, allowing calculation of the peak luminosity using the observed rate of decline in luminosity. As the peak luminosity does not vary greatly, about 0.3 magnitudes between different type Ia supernovae, these are used as "standard candles" (sources for which their absolute brightness can be determined based on the type of object). In other words, as the absolute magnitude of type Ia supernovae has little spread, the distance can be determined using the apparent magnitude and the distance modulus, therefore serving as important distance markers for distant galaxies. [12, pp. 680–686]

  The exact mechanism that causes a type Ia supernova is still unclear, but it is thought that these are caused by the destruction of white dwarfs in a close binary system. If, in such system, the white dwarf accretes enough matter from its companion so that its mass reaches or gets close to the Chandrasekhar limit (the maximum mass a white dwarf can reach before the relativistic degeneracy pressure fails to balance gravity), a thermonuclear ignition of the carbon core occurs resulting in a catastrophic explosion. Unlike dwarf and classical novae, a type Ia supernova is thought to destroy the white dwarf and leave no stellar remnants [12, pp. 686–689] [29, pp. 87, 104].

Cataclysmic variables produce their $H\alpha$ emission lines via their accretion disk discussed in section 1.1.1.

Symbiotic Binaries

Symbiotic binaries are similar to cataclysmic variables in that the accreting component is often a white dwarf. However, in a symbiotic binary the companion star is a red giant and does not lose matter due to an overflowing Roche lobe; instead matter is lost due to its stellar wind. This matter then accretes onto the primary component via an accretion disk, which produces the $H\alpha$ emission lines, as for CVs. [60][12, p. 673].

X-ray Binaries

In these systems the accreting component is either a neutron star or black hole, accreting matter from a non-degenerate companion star. For low-mass X-ray binary system the companion star is a late type main-sequence star, as in CVs. A a high-mass binary has a red giant as its companion, same as in a symbiotic binary. These systems have been observed with luminosities close to the Eddington limit, which can be used to determine the temperature of the in-falling matter at the surface/event horizon of the accreting object. This temperature, about $2 \times 10^7 K$, corresponds to where blackbody radiation peaks in the X-ray, hence explaining the large X-ray emission of these systems. As with other binary systems, the matter is accreted via an accretion disk which produces the observed $H\alpha$ emission. [12, p. 672][15, pp. 145–149]

1.2.2 Young Stellar Objects

Young stellar objects are stars in their early stages of evolution with no hydrogen burning. The earliest evolution stage of a star is the *protostar*, which occurs after cloud collapse and fragmentation when the increase in density makes the gas opaque. As a result the radiation inside the cloud core is trapped, resulting in an increase in temperature and therefore gas pressure, with the cloud core coming into hydrostatic equilibrium.

This is followed by a period of spherical accretion, with an accretion disk forming in the later stages of accretion due to the angular momentum of the in-falling matter. The accretion disk emits mostly at the far-infrared and sub-mm wavelength; and produces more luminosity than the central protostar. Once an accretion disk is formed jets or bipolar outflows, are also commonly observed and these are called Herbig-Haro objects. Depending on the mass of the of the protostar it then either emerges on the pre-main sequence as a classical T Tauri star (TTS) ($< 2M_\odot$) or a Herbig Ae/Be star ($2 - 10M_\odot$). T Tauri stars (TTS) can be split into two groups, classical TTS and weak-line TTS. Classical TTS are in an early stage of pre-main sequence evolution and still have an optically thick accretion disk, which produces excess emission at the IR wavelengths. These also often exhibit strong hydrogen (Balmer series), Ca II and iron emission. Weak TTS are further along the pre-main sequence and are characterised by an optically thin accretion disk with much weaker emission lines. Herbig Ae/Be stars are the analogy of TTS for more massive protostars, exhibiting many of the same characteristics as TTS such as emission at the Balmer lines and excess infrared radiation. However unlike TTSS, this IR radiation is due to generally being surrounded by some remaining dust and gas. Herbig Ae/Be stars have not been studied as well as TTS, mainly due to their much shorter lifetime on the pre-main sequence and hence smaller known sample. [44, pp. 123–134] [12, pp. 433–441].

1.2.3 M Dwarfs or Red Dwarfs

M dwarfs or red dwarfs are main sequence stars of spectral type M. They are some of the coolest and least massive stars, residing in the bottom right of the main sequence on a colour-magnitude diagram. They are also the most numerous stars in our Galaxy. A subgroup of these exhibit chromospheric $H\alpha$ emission, which has been shown to be coupled with the rotation of the star [36, 49].

1.2.4 Be Stars

Be stars are rapidly rotating O, B and A-type stars with luminosity classes V-III, which at some point exhibited emission at one or more of the Balmer series. Their emission lines are thought to be from their equatorial, circumstellar disks fed by gas ejected from the surface of the star due to their large rotation rate. These emission lines are not always observed and there is currently no comprehensive explanation for the on/off switch of the excess emission and it is currently not known how a B star becomes a Be star [17].

1.3 Photometric Surveys

Photometry measures the photon count for a specific pass band filter (or no filter), which only lets through photons of a certain wavelength range while blocking all others. This then gives the photon count for the different wavelength filters used; for large scale surveys this is applied to fields of view to get observations for a large number of objects at once, providing an effective way to get colour and apparent magnitude observations for many objects. Photometric surveys that include the parallax (i.e. distance), such as Gaia, give a colour-magnitude diagram (CMD). This can then be used to get a general idea about an object of interest based on its position on the CMD. However, in order to get a more complete picture, the spectrum has to observed using spectroscopy, which splits the incident light using a prism or diffraction grating; this is far more expensive and time intensive than photometry.

As covered above, a large number of pre- and post-main sequence type objects exhibit excess $H\alpha$ emission; many of these object types are of particular interest to the astronomical community, as they help to confirm and improve models of formation and evolution of these systems. Some of these only have a limited number of confirmed sources, hence large scale photometric $H\alpha$ surveys are important to expand the sample of potential sources. A large-scale photometric $H\alpha$ survey, using a narrow band $H\alpha$ filters, such as the $H\alpha$ Survey of the Northern Galactic Plane

(IPHAS, section 3.1), provides the opportunity to find many $H\alpha$ emitting sources at once; these can then be followed up spectroscopically for sources of interest to identify the exact type and undertake further studies.

Combining a large scale photometric survey, that includes the absolute magnitude (and therefore allows accurate placements of the sources on a CMD), with a large scale $H\alpha$ survey can then be used to efficiently find $H\alpha$-emitting sources. The advantage of combining an $H\alpha$ survey, such as IPHAS, with a survey such as Gaia, is that it gives the absolute magnitude; this combination can be used to identify $H\alpha$ emitters that would otherwise not be identifiable due to reddening. For the Witham et al. (2008) [66] $H\alpha$ emitter catalogue, discussed in more detail in section 1.4, emitters were selected using colour-colour plots of r-i and r-$H\alpha$. However, using only these two dimensions means that r-$H\alpha$ emitters that sit in the lower reddened main-sequence and giant branch are missed, as they show up as non-emission line objects in the unreddened main-sequence branch. A colour-colour plot is shown in in the lower part of figure 3.2 with the unreddened main-sequence branch at the top and the reddened main-sequence and giant branch below.

This report describes the steps and results of producing a catalogue of $H\alpha$ emitters and potential emitters using a crossmatched Gaia/IPHAS catalogue and unsupervised machine learning techniques. Chapter 2 gives a brief introduction to machine learning and its uses in astronomy and astrophysics, followed by a more detailed explanation of the clustering algorithms used. Methods and data used are covered in chapter 3, with the selection of $H\alpha$ emitters discussed in chapter 4. To ensure that the results are accurate, several approaches of validation are covered in chapter 5, followed by results and discussion in chapter 6.

1.4 Similar Work - Witham et al. Catalogue

The IPHAS catalogue of $H\alpha$ emission line sources in the northern galactic plane produced by Witham et al. (2008) [66] is a catalog of 4853 point sources with a high likelihood of being $H\alpha$ emitters; identified from the initial data release of the IPHAS survey. The main differences between the Witham et al. (2008) [66] catalogue compared to the one presented here is

- it uses a different approach to identify $H\alpha$ emitters, briefly explained in section 1.4.1, or for full details see Witham et al. (2006; 2008) [65, 66].
- it uses the full initial IPHAS dataset [21], whereas this catalogue is based upon a cross-match between Gaia DR2 and IPHAS DR2. It is also worth noting that some of the values in the initial IPHAS data release were re-calibrated in DR2. [6]

A comparison between the Witham et al. and this catalogue is done in section 5.4.

1.4.1 Selection Algorithm

This is a brief summary of the selection algorithm used by Witham et al. (2008) [66] to create their $H\alpha$ emitter catalogue; the full description can be found in [65].

Initial data cuts were applied to the IPHAS fields based on the seeing in each band and the average stellar ellipticity. Further cuts were also applied to individual sources, covered in section 3.1 of Witham et al. (2008) [66].

The following selection algorithm was then applied to the remaining 12959 IPHAS fields. The sources were split into four magnitude bins; and for each field and magnitude bin combination, r-$H\alpha$ and r-i plots were created. On each of these plots an initial straight line least-squares fit is performed as an initial attempt to fit the main stellar locus. This works well for low density fields. However, in fields with a higher density, the stellar locus is often split due to different stellar loci and different reddening, which results in the upper locus often being above the fit. In this scenario, using the initial fit to identify $H\alpha$ emitters would result in incorrect selection, therefore four iterations of σ-clipping are performed to force the fit to follow the upper boundary of the main stellar locus. In some cases the final fit is not appropriate, such as in fields where the stellar locus is not split. In those cases the initial fit is used for the selection of $H\alpha$ emitters; otherwise the final fit is used. With the appropriate fit determined, objects significantly above the fit were identified as $H\alpha$ emitters. Figure 1.1 shows the colour-colour plots for the four magnitude bins for IPHAS field 2373 along with the fits and selection cuts.

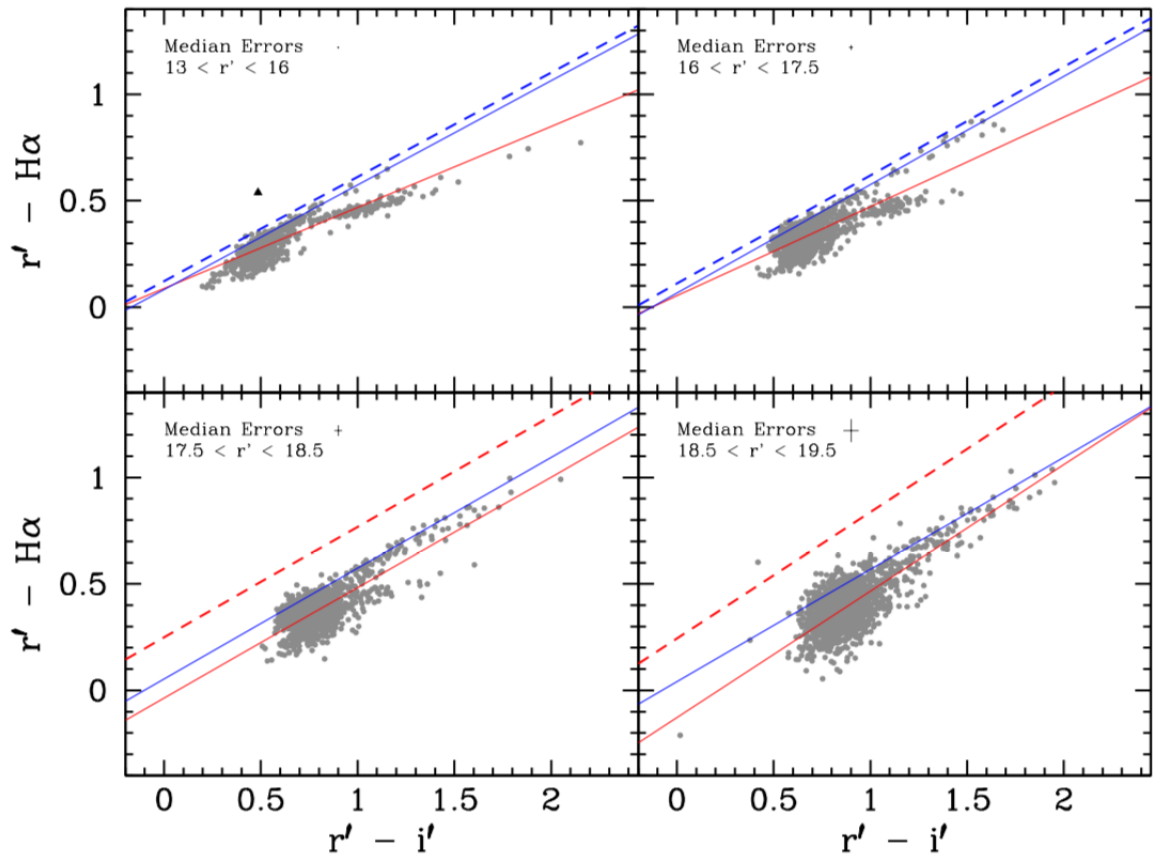


Figure 1.1: The solid red line is the initial least-squares fit, with the solid blue line representing the final fit to the upper locus from the iterative σ-clipping. The dashed line is the actual cut used to select $H\alpha$ emitters, shown in red if the initial fit is used, otherwise blue. However, these cuts are only approximate, as actual selection also considers the errors on individual objects. $H\alpha$ emitters are shown as triangles. Figure is from Witham et al. (2008) [66].

Chapter 2

Machine learning

2.1 Machine Learning

The large increases in data collection in the last ∼10 years has changed our everyday life dramatically. This effect can also be seen in astronomy and astrophysics, with the amount of data collected increasing at an unprecedented rate. Some examples include the Hubble Space Telescope, producing ∼3 GB of data per day; its newer replacement, the James Webb Space Telescope, expected to produce ∼57.5 GB per day [7]; and on the extreme end of the scale, the Square Kilometre Array is set to produce 160 TB of data per second, which is 14×10^9 GB per day. A more recently released dataset, Gaia DR2, contained the magnitudes of 1.69 billion sources, with parallaxes and proper motion for ∼ 1.3 billion of those, which is an increase in the number of sources by a factor of ∼700 compared to its predecessor mission Hipparcos.

Analysing and processing this amount of data using traditional approaches is becoming more and more difficult, requiring more automated approaches to group, structure and classify the data. This is where the recent improvements in machine learning, made possible by advances in computer processing power, provides an opportunity to make new discoveries and analyse more data at a larger scale.

Machine learning offers powerful classification algorithms that allow classifying data based on models that were trained on past labelled data (supervised machine learning). Labelled data is data that has already been classified and is used to "teach" the model using a training algorithm. However, in order to train a model that generalises well a large amount of high quality training data is required, with the amount depending the complexity of the task. Creating a suitable training dataset is often only achievable by extensive manual work; projects such as Galaxy Zoo, a citizen science project for morphological classification of 304,122 galaxies (Galaxy Zoo 2) [63], are a good way to achieve this. It does, however, require several years to complete and the participation of the public. Another example of a citizen science project is Gravity Spy, which

involves classifying instrumental and environmental sources of noise to allow the training of a model that can then detect these *glitches* in much larger datasets and therefore improve LIGO's sensitivity. [68]

Unsupervised machine learning on the other hand, does not require a model to be trained on existing labelled data. Instead it is used to identify structures, groups and patterns in unprocessed and unstructured data to either make future analysis easier, identify new features or discover new relationships in the data. In addition it is also commonly used for dimensionality reduction of high dimensionality datasets.

2.1.1 So what is Machine Learning?

Machine learning (ML) is a subfield of Artificial Intelligence concerned with algorithms that allow computers to *learn* from data. The two common definitions of ML below further specify this concept. The earliest definition of machine learning is from Arthur Samual in 1959, who defined it as "the field of study that gives computers the abilities to learn without being explicitly programmed" [50]. A more formal definition was given by Tom M. Mitchell - "A computer is said to learn from experience E with respect to some class of task T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" [33].

A simple example of ML in most peoples everyday life is an email spam filter. As a human determining whether an email is spam or not is a simple task, done by a briefly viewing the email and looking for common spam patterns. Machine learning allows training of a model that is able to identify these patterns from a large amount emails that have already been classified as either spam or non-spam emails. This model can then be used to classify future emails as either spam or non-spam. Other common examples of machine learning are web search engines such as Google, targeted advertisement, the movie/show recommendation system by providers such as Netflix, and many others.

Machine learning is often split into three different types: supervised machine learning; unsupervised machine learning; and reinforcement learning. More details for the first two is given in sections 2.2 and 2.3 respectively. Reinforcement learning is concerned with training a system (or agent) that improves its performance based on interaction with the environment. This is not overly relevant for most applications in astronomy and astrophysics, a more detailed explanation of this is therefore not included here. It is also worth pointing out that these are just rough categories and there are many other ways in which to differentiate machine learning algorithms, such as whether or not they can learn incrementally on the fly (online vs batch learning), or whether they make predictions by comparing the data to the training data or instead build a

predictive model from the training data (instance-based versus model-based learning).

2.2 Supervised Machine Learning

As already briefly discussed above, supervised machine learning is used to predict or estimate a target variable from input data. Input data is made up of the features used to predict the target variable; for a simple example such as predicting house prices, the features used as input data could be the size of the house, number of bedrooms and bathrooms, age of the house and anything else that might affect the house price. It is important to note that the input data for training and for making predictions/estimations has to consist of the same features.
Figure 2.1 schematically shows how a model is trained using a machine learning algorithm and labelled data; the trained model can then be used to make predictions of the target variable. Labelled data consists of the input data (i.e. the features) and their associated "true" values of the target variable. Sometimes the labelled data is also called training data, but this can be misleading as the labelled data is often split for the training and testing of the model. In this case the data used for training and testing are called training and testing data respectively, although both are labelled data. Supervised machine learning is often split into two types: regression and classification. Regression is the prediction of a continuous output variable, such as house prices, whereas classification is concerned with output variables that are discrete, for example an e-mail spam filter.

A classic introductory classification example is the *Iris* dataset, containing the measurements

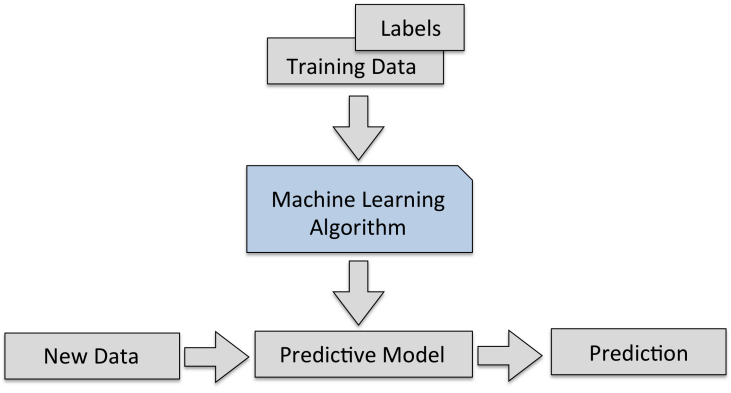
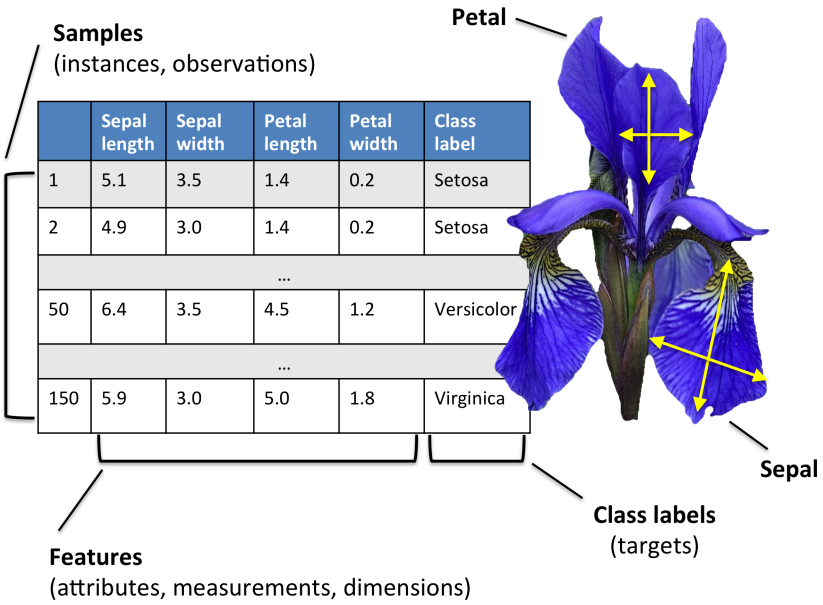


Figure 2.1: Schematic visualisation of how a supervised machine learning model is trained and used for estimation/prediction. Figure from [45]

of three different species (*Setosa, Versicolor* and *Virginica*) of the iris flower. Figure 2.2 shows three samples of the dataset, one for each of the species, i.e. the target variable in this case.

The decision boundary, for two of the four features, of a support vector machine (SVM) model



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

Figure 2.2: Table showing three labelled data entries of the *Iris* dataset, one for each of the possible classifications. Each entry is made up of the values for each of the four features and its assigned label (i.e. value of the target variable). Figure from [45].

trained on the *Iris* dataset is shown in figure 2.3. The plot visualises how the model splits the feature space into the different classes of the target variable; the values of labelled data and their "true" classification are also shown, providing a visualisation of how well the model is performing. In this instance the model performs very well, as there are only a small number of misclassifications at the green and blue boundary.

2.3 Unsupervised Machine Learning

Unsupervised machine learning is not used to predict a target variable; instead it is used to identify structures, patterns and groups in the data. Hence it does not require labelled data, as there is no target variable. A common application of unsupervised machine learning is the clustering of data, i.e. grouping data points based on their spatial locations in the feature space. An example of clustering algorithms is k-means, illustrated in figure 2.4. K-means clustering groups n samples into k partitions, where the number of partitions, k, is a parameter of the algorithm and has to be known beforehand or determined by trial and error. However, as k-

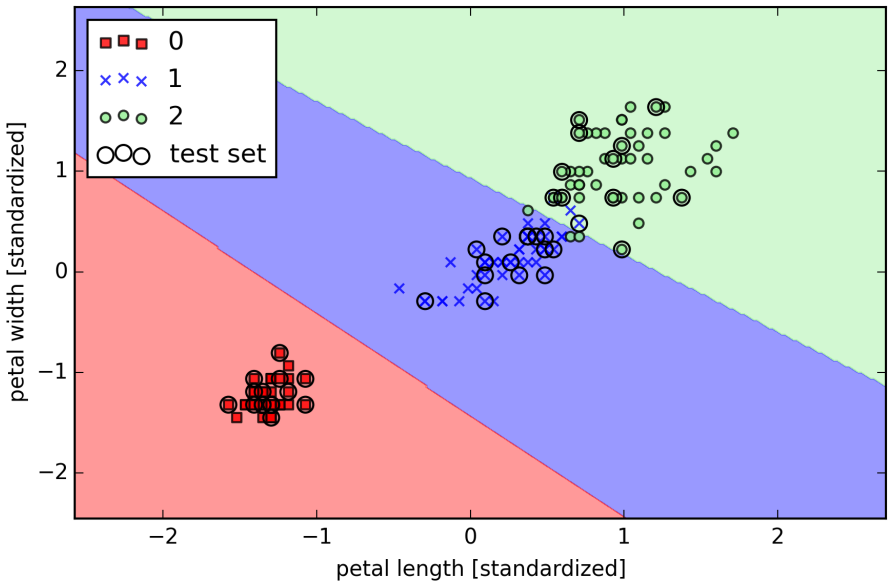


Figure 2.3: The decision boundary of a SVM model trained on the *Iris* dataset to predict the different species. Points with a circle around them were not used for the training of the model and are commonly referred to as the test set. This decision boundary shows two of the four features of the dataset and therefore does not give a complete picture, but still shows quite clearly that the model is performing well, and there are only a very small number of misclassifications at the blue-green boundary. Figure from [45].

.

means clustering works based on distances to the central points of the cluster, called centroid points, the shape of the clusters is defined by the distance function used (e.g. spherical for euclidean distance) and hence does not support arbitrary shaped clusters.

K-means is an example of a partitioning algorithm. A different type of clustering is called density-based clustering, which uses the density of data points to identify the clusters. Some algorithms are able to identify clusters of different densities and others form clusters based on density specified by parameters. An example of the latter is the DBSCAN algorithm, which stands for "Density-based spatial clustering of applications with noise". One of the advantages of this algorithm compared to k-means is that it is able to identify clusters of arbitrary shape and differentiate between noise and cluster points; in k-means every data point is assigned to a cluster. The DBSCAN algorithm is explained in more detail in section 2.6.2.

2.4 Semi-Supervised Learning

Supervised and unsupervised learning are the two traditional main types of machine learning, with semi-supervised machine learning sitting between the two. In supervised machine learning, only data that has associated labels for the target variable can be used, which is often expensive

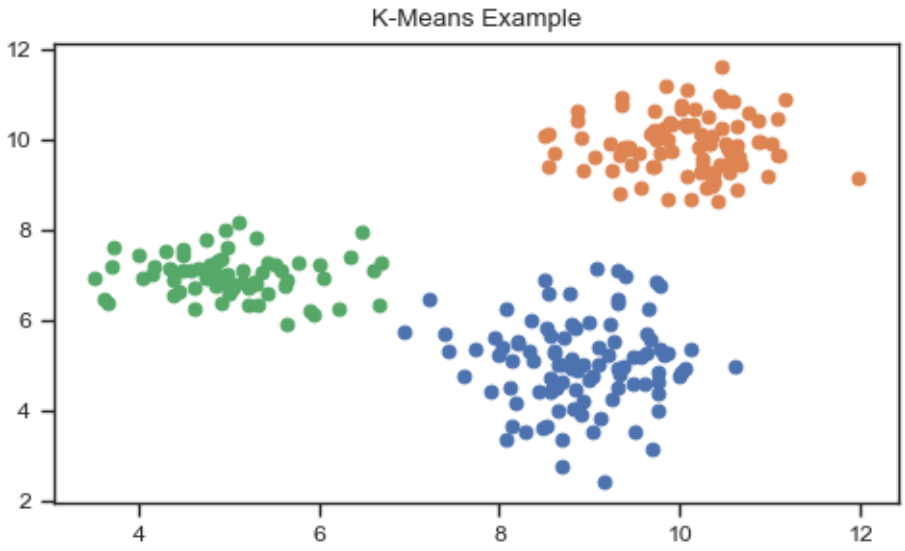


Figure 2.4: K-means applied to an example dataset consisting of data points drawn from three different normal distributions. Worth highlighting are the spherical shape of the clusters and the fact that all points are assigned to a cluster even if they are in much lower density regions and would not necessarily be considered as part of a cluster by a human.

in terms of both cost and time to acquire, unlabelled data, on the other hand, is generally much easier to acquire, but cannot be used for supervised learning.

The standard setting of semi-supervised machine learning is a supervised machine learning approach, with the unlabelled data used to provide additional information on the distribution of the features. In other words, the performance of a supervised machine learning model can often be improved by using unlabelled data to provide extra information on the feature space, that is not captured by the, often limited, labelled data. In this setting the dataset is split into two parts, the samples X_l for which labels Y_l are available and the samples X_u for which no labels are available.

Another approach is to view semi-supervised learning as unsupervised learning guided by constraints from the labelled data. For more details on semi-supervised machine learning see Chapelle (2006) [13].

Semi-supervised machine learning is of specific interest for areas in which large quantities of unlabelled data exists, and only limited amounts of labelled data; such as astronomy and astrophysics. Labelled data exists for many problems in databases such as SIMBAD [58]. However, with the amount of data collected increasing rapidly due to large scale surveys such as Gaia (1.6 billion sources), semi-supervised learning can provide a way to use these large quantities of new data with much less manual effort. An example of semi-supervised machine learning applied to supernova classification is briefly covered in section 2.5.

2.5 Machine Learning Examples in Astronomy

This section introduces some interesting applications of machine learning in the area of astrophysics.

2.5.1 Semi-supervised learning for photometric supernova classification

This paper/study [48] uses semi-supervised machine learning for supernovae type classification. All of the available data (i.e. unlabelled and labelled) is used to create a lower dimensional representation and, using this, a standard supervised machine-learning classifier is trained on high-quality training data (labelled data).

The advantage of this approach, compared to using only the labelled data to train a supervised machine learning model, is that it utilises all of the available data to create a more complete low dimensional representation of the feature space. This meant that the authors did not have to estimate parameters such as redshift, stretch or reddening, as the variations appear as gradual variations in the low-dimensional space; this works as the observed data is collected at high resolution over the variations of these parameters.

Reducing the data to a lower-dimensional feature space was done using a diffusion map, which is a non-linear reduction technique giving a lower dimensional space in which the Euclidean distance between any two points approximates the diffusion distance; a distance measure that captures the intrinsic geometry of the data set [47, 48]. A random forest (an averaged ensemble of randomly varied decision trees) is then used to train a classifier in the low-dimensional feature space to classify the supernovae. The resulting classification model was entered in the SN Classification Challenge, showing that it is competitive with the other entrants. However, due to the limited training set this classifier can only be used for redshifts that are available in the labelled training data. To further improve this, larger/deeper labelled training sets are required. For the full details see Richards et al. (2009; 2011) [47, 48].

2.5.2 Generative Adversarial Networks recover features in astrophysical images of galaxies beyond the deconvolution limit

Images taken from a ground telescope are limited by various sources of noise, such as the detector, atmosphere, and the sky background. The blurring introduced by the combination of telescope and atmosphere is described by the point spread function (PSF). The convolution of the true light distribution with the PSF (plus other sources of noise) gives the image taken by the

telescope. The reverse process, deconvolution, to remove the effects of the PSF, and retrieve the true image is limited by the Shannon-Nyquist sampling theorem [56, 40]. A standard approach for an inverse problem such as this, is using knowledge (priors) from forward modelling to allow the algorithm to make a more informative decision when choosing from the possible solutions.
This paper [52] shows that machine learning techniques can go beyond the deconvolution limit by building effective priors from high-quality training data. Generative Adversarial Networks (GANs) introduced by Ian Goodfellow et al. (2014) [22] use two competing networks: one to estimate the underlying distribution (the Generator); and a competing network (the Discriminator) that estimates the probability of the sample coming from the training data rather than the Generator. In this paper, GANs are used for image-to-image translation and are schematically shown in figure 2.5. Artificial noise is added to the original image, and is feed through the Generator, which attempts to recover the original image. Only in the training phase does the Discriminator try to distinguish the recovered image from the original image. These two networks are then trained together by competing against each other.
It is worth pointing out that the Discriminator network is only used in the training phase, not during testing or actual prediction/estimation. The results were evaluated quantitatively using the Peak Signal Noise Ratio, which is the ratio between the maximum possible power of a signal (original image) and the power of the noise (difference between recovered and original image). Qualitative assessment was done by comparing the original image, the degraded image, the recovered imaged and the deconvolved image.
This comparison showed that this approach far exceeds traditional deconvolution techniques, but is limited by the training set, as the conditions of the training data (redshift, camera etc.) and the actual data have to be in similar conditions. For full details see Schawinski et al. (2017) [52].

Some further works using machine learning are *Machine Learning Classification of Gaia Data Release 2* [5] and *A Deep Learning Approach to Galaxy Cluster X-ray Masses* [39], which constructs and tests a convolutional neural network (commonly used for image recognition), to estimate mass values of galaxy clusters using X-ray mock observations. There are many other applications of machine learning in astronomy and astrophysics.

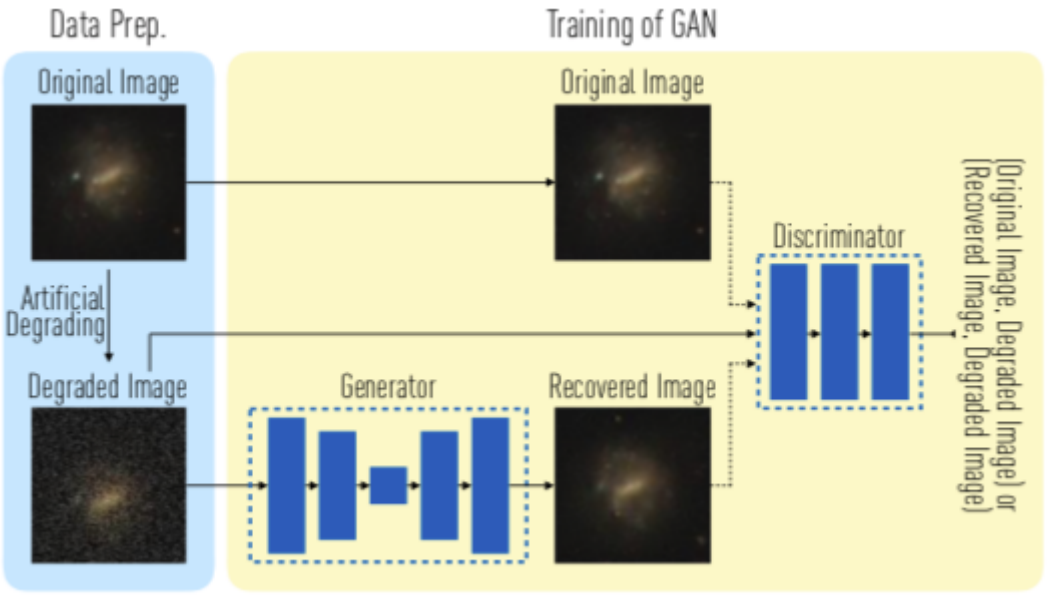


Figure 2.5: A schematic visualisation of the generative adversarial networks used in galaxy image deconvolution, discussed in section 2.5. Image from [52].

2.6 Relevant Machine Learning Algorithms in Detail

2.6.1 k-Nearest-Neighbours

The k-nearest-neighbours (k-NN, kNN or NN) algorithm is an instance based classification or regression machine learning technique. Instead of fitting a model to the training data, this algorithm selects the k-nearest-neighbours (of the labelled data) for a given input sample, and then performs a majority vote using the labels of the k-nearest-neighbours. For classification, the most common class among the k-nearest-neighbours is then assigned to the new input sample and, for regression, the output is the average (of the k-nearest-neighbours). The only parameter of this algorithm is the number of nearest neighbours to use, k. Figure 2.6 shows an example of kNN classification for different values of k.

2.6.2 DBSCAN

As DBSCAN is the clustering algorithm used as part of the $H\alpha$ emitters selection process, this section gives a brief summary of the DBSCAN algorithm. For the full details of the algorithm, see the original paper by Martin Ester et al. (1996) [19].

The key idea of the DBSCAN algorithm is that the neighbourhood of each cluster contains at least a minimum number of points, denoted $MinPts$. The neighbourhood, N_ϵ, of a point is defined as all points within a certain radius, ϵ. The distance function, $dist(p, q)$, for two points p and q, determines the shape of the neighbourhood. For example the Euclidean distance gives a

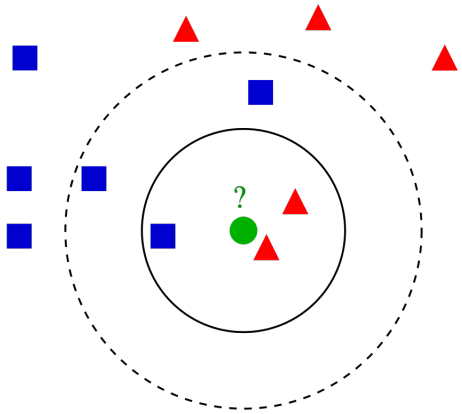


Figure 2.6: kNN classification for different k values. The solid circle shows kNN for $k = 3$, which results in the new point in question to be classified as a red triangle. A value of $k = 5$, represented by a dashed circle, results in the new point being classified as a blue square. Image credit to [10].

spherical neighbourhood in 2D space, whereas the Manhattan distance (the sum of the absolute differences of the Cartesian coordinates) gives a rectangular neighbourhood. The remainder of this section will assume use of the Euclidean distance function, although other distance functions are also completely valid.

The DBSCAN algorithm differentiates between two types of cluster points: points inside the cluster, called core points; and points on the edges of the cluster, called border points. This makes sense as cluster will generally have a lower density on their borders/edges compared to their innermost parts. Therefore border points will in general have considerably less points in their neighbourhood compared to core points. Hence a lower $MinPts$ value (than for the core part of the cluster) would be required to include the complete cluster. However, this may lead to the inclusion of noise points as part of the cluster. By differentiating between these two types of cluster points, the DBSCAN algorithm is able handle the different criteria without requiring an extra parameter.

A core point satisfies the core-point condition, $|N_\epsilon| \geq MinPts$, whereas a border point does not satisfy the core-point condition, but is considered a cluster point as it is in the neighbourhood of a core point [19]. In other words, if point p is a core point (meets the core condition) and its neighbourhood, N_ϵ, contains the point q, which does not meet the core condition, then point q is considered a border point and is part of the same cluster as point p.

In order to define a cluster, the concepts of *directly density reachable*, *density reachable* and *density connected* have to be defined.

A point p is *directly density reachable* from a point q if q is a core point and p is in the neigh-

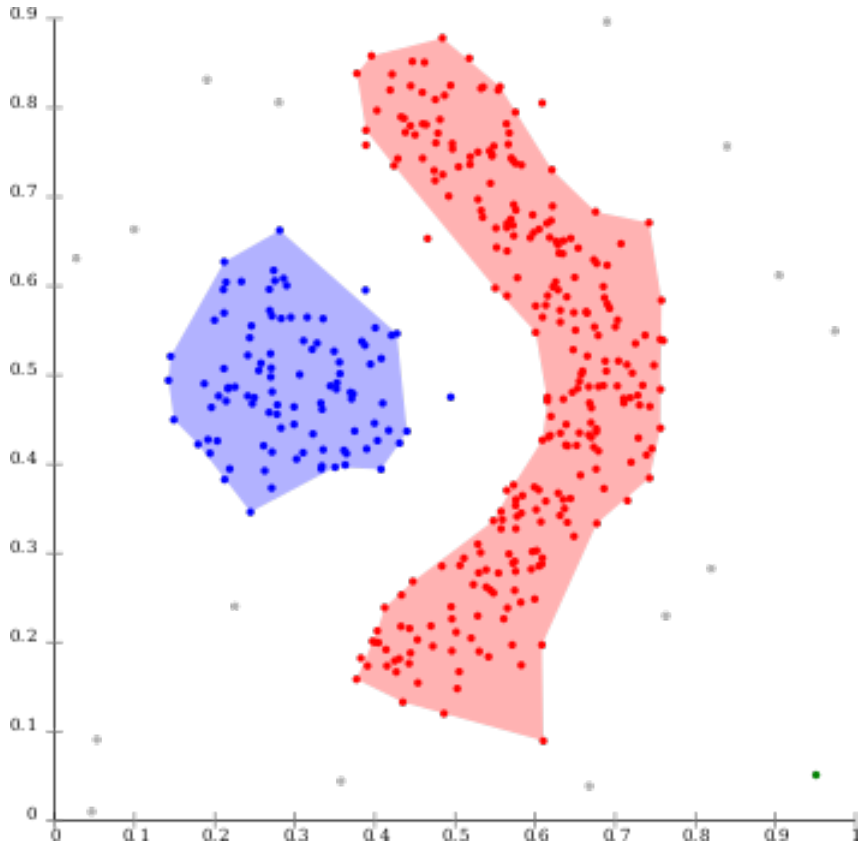


Figure 2.7: Example of DBSCAN clustering. Noteworthy are DBSCANs ability to determine the number of clusters on its own (i.e. number of clusters is not a parameter that has to be specified) and identifying clusters of arbitrary shape; compared to an algorithm like k-means which requires the number of clusters as a parameter and selects clusters in a spherical shape. Image from [11].

bourhood of q, as shown in figure 2.8.

Density reachable is the canonical extension of *directly density reachable*, shown in figure 2.9.

A point p is *density connected* to a point q if there is a point o such that both p and q are *density reachable* from o, shown in figure 2.10.

A cluster, C for a dataset, D, can then be defined as:

- $\forall p, q$: if $p \in C$ and q is density reachable from p (Maximality)
- $\forall p, q \in C : p$ is density connected to q (Connectivity)

In other words a cluster is defined as all core points that are *density reachable* plus all points that are within the neighbourhood of one these core points.

Identifying a cluster can be done in a step-by-step approach:

1. Select an arbitrary point, p in the database.
2. Retrieve all points that are density reachable from this seed point.

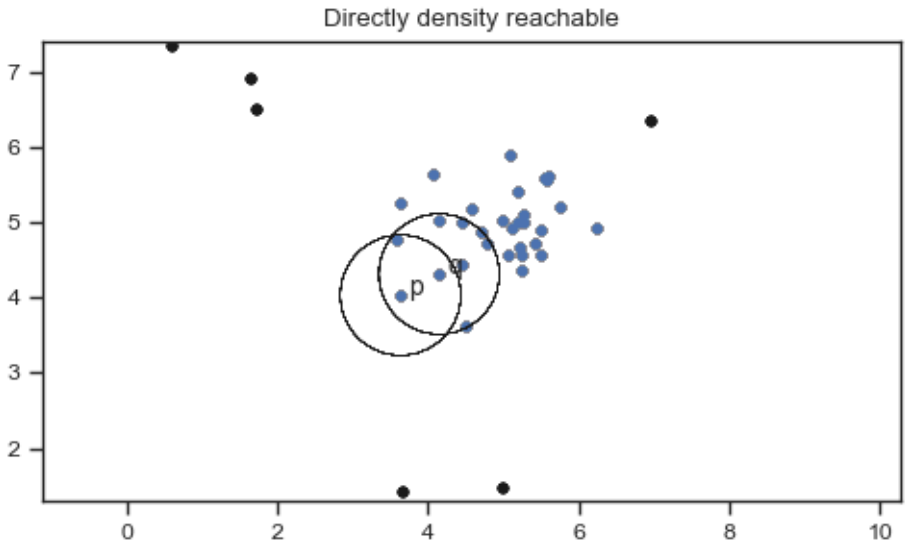


Figure 2.8: Plot of example data with the concept of *directly density reachable* shown for the two points q and p, where q is the core point and p is in the neighbourhood of q.

3. If p is a core point (i.e. meets the core point condition), then this results in a cluster.

4. If p is a border point, no points are density reachable from p and the next point in the dataset is visited.

5. Clusters with a minimum distance ($dist(S_1, S_2) = mindist(p, q)|p \in S_1, q \in S_2$) apart are merged.

The steps are then repeated over the whole dataset. It is worth noting that border points can be taken from one cluster and transferred to a new cluster, i.e. if a cluster is being created and a point, p, is a border point of an existing cluster, then if p is directly density reachable from one of the core points in the new cluster, it is "transferred" to the new cluster.

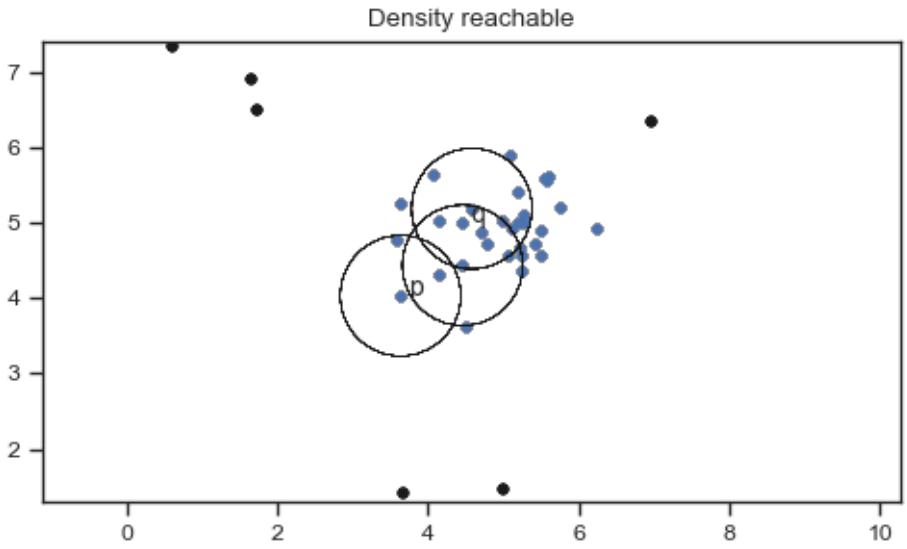


Figure 2.9: Plot of example data with the concept of *density reachable* shown for the points q and p. Point p is *density reachable* from q as it is *directly density reachable* from a point that is *directly density reachable* from point q. This can be chained over many points.

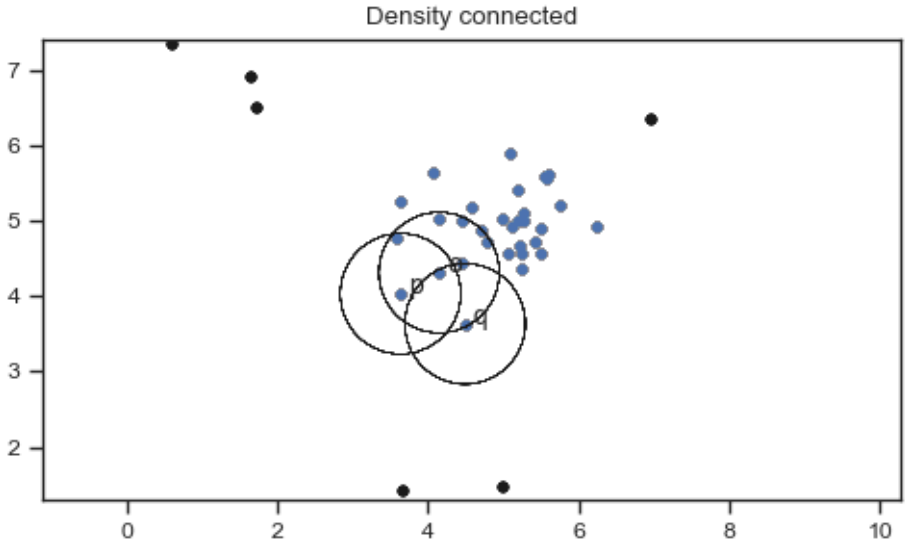


Figure 2.10: Plot of example data with the concept of *density connected* shown. Point o is a core point and connects both points p and q as both are *density reachable* from o.

Chapter 3

Methods

3.1 Gaia/IPHAS Catalogue

The catalogue of $H\alpha$ emitters presented here is based on the Gaia/IPHAS value added catalogue (VAC) produced by Simone Scaringi et al. (2018) [51]. This section gives a brief overview of the Gaia/IPHAS catalogue, but for the full details on the catalog, see Scaringi et al. (2018) [51].
The second data release of the Gaia mission provides G-band photometry for approximately 1.7 billion sources; and G_BP (330-680 nm) and G_RP (630-1050 nm) band photometry and parallax measurements for 1.3 billion sources, which allows calculation of distances and absolute magnitudes [9]. It is worth noting that the absolute magnitude in the the Gaia/IPHAS VAC are not corrected for extinction and are therefore upper limits.
The Isaac Newton Telescope (INT) Photometric Hα Survey of the Northern Galactic Plane (IPHAS) is a large-scale survey of the northern Milky Way over the latitude and longitude ranges $-5^\circ < b < +5^\circ$ and $30^\circ < l < 215^\circ$, respectively. The data is collected in the r and i broad-band filters, along with the $H\alpha$ narrow-band filter using the Wide Field Camera (WFC) on the 2.5m Isaac Newton Telescope in La Palma. Full details can be found in Drew et al. (2005) [18] and Barentsen et al. (2014) [6].
To produce the Gaia/IPHAS VAC the following cuts were performed on the Gaia and IPHAS catalogue before crossmatching. For IPHAS DR2, objects had to meet the following criteria:

- measurements in all three bands (r, i, $H\alpha$);
- did not exceed the saturation limit in any of the bands ($r > 13$, $i > 12$, $H\alpha > 12.5$);
- photometric errors <= 0.1 in all bands; and
- not flagged as blended or affected by bright neighbours in any band.

and for Gaia DR2

- G-band flux S/N (phot_g_mean_flux_over_error) > 5;
- parallax S/N (parallax_over_error) > 5; and
- within the area $20° < l < 220°$ and $-6° < b < 6°$ (slightly larger than the IPHAS area).

The crossmatching of the catalogues also took into account that all Gaia DR2 sources use epoch 2015.5, whereas IPHAS DR2 uses the epoch of observation, i.e. anytime between 2003 and 2012. This was done by winding back the proper motions of the Gaia DR2 objects to match the epochs of IPHAS DR2, ensuring high proper motion objects were correctly cross-matched. For the full details on the cross-matching method see Scaringi et al. (2018) [51] section 2.

The resulting catalog contains 7,927,224 sources and defines the two quality parameters f_c and f_{FP}, which were added to clean the Gaia/IPHAS VAC from sources with unreliable parallax measurements.

The f_c parameter is a measure of how good or bad the astronomic fit of a specific source is compared to other sources within a similar apparent magnitude bin. To calculate f_c all sources were binned in m_G with a bin width of 0.1 and each source was then assigned a percentile based on the its reduced χ^2_v value. As this parameter allows us to remove the fraction of sources with a bad astrometric fit, it is called the "completeness fraction", f_c.

Gaia DR 2 contains some spurious parallax measurements (either very large or negative) as discussed in Lindegren et al. (2018) [27]; these are expected to be removed in future data releases from Gaia. Scaringi et al. (2018) [51] create a "mirror sample" of Gaia sources that are known to have bad parallax measurements and perform the same crossmatch as with the actual catalogue. This provides a dataset with a confirmed bad astrometric fit, while still retaining the statistical properties of the actual catalogue. All sources, including the mirror sample, are again binned into 0.1 m_G bins; each bin is then sorted in increasing order by χ^2_v and binned further into blocks of 1000 sources. Each source is then assigned a false-positive fraction, f_{FP}, with respect to its block; in other words, for a given source $f_{FP} = \frac{N_{neg}}{N_{pos}+N_{neg}}$, where N_{neg} is the number of "mirror sample" sources in the block and N_{pos} the number of sources from the actual catalogue. The assumption made for this quality parameter is that the Gaia DR2 processing produces some spurious astrometry, resulting in positive or negative parallaxes. This assumption is tested by calculating the absolute G-band magnitude for the "mirror sample" using the absolute value of the parallax and plotting this on the CMD, which showed that this "mirror sample" sits remarkably well in the region corresponding to unreliable targets. For full details on these two quality parameters along with suggested cuts for clean CMD data see section 3 in Scaringi et al. (2018) [51].

The following data cuts were used: $f_c < 0.98$ and $FP < 0.02$, as per the suggestion in the paper.

Additional saturation cuts were also performed, dropping all sources with $r < 13.5$, $i < 12.5$ and $H\alpha < 13$. These cuts were altered from the initial cuts, $r < 13$, $i < 12$ and $H\alpha < 12.5$ based on Barentsen et al. [6], after it was discovered that a large number of the brightest sources selected as $H\alpha$ emitters, were actually not emitters when checked against their respective LAMOST spectrum. For more details see section 5.2. This left 7,373,236 sources, with the differences between the cleaned and original data shown in figure 3.1.

From the available colours and magnitudes, the two colours, r-i (IPHAS) and r-$H\alpha$ from

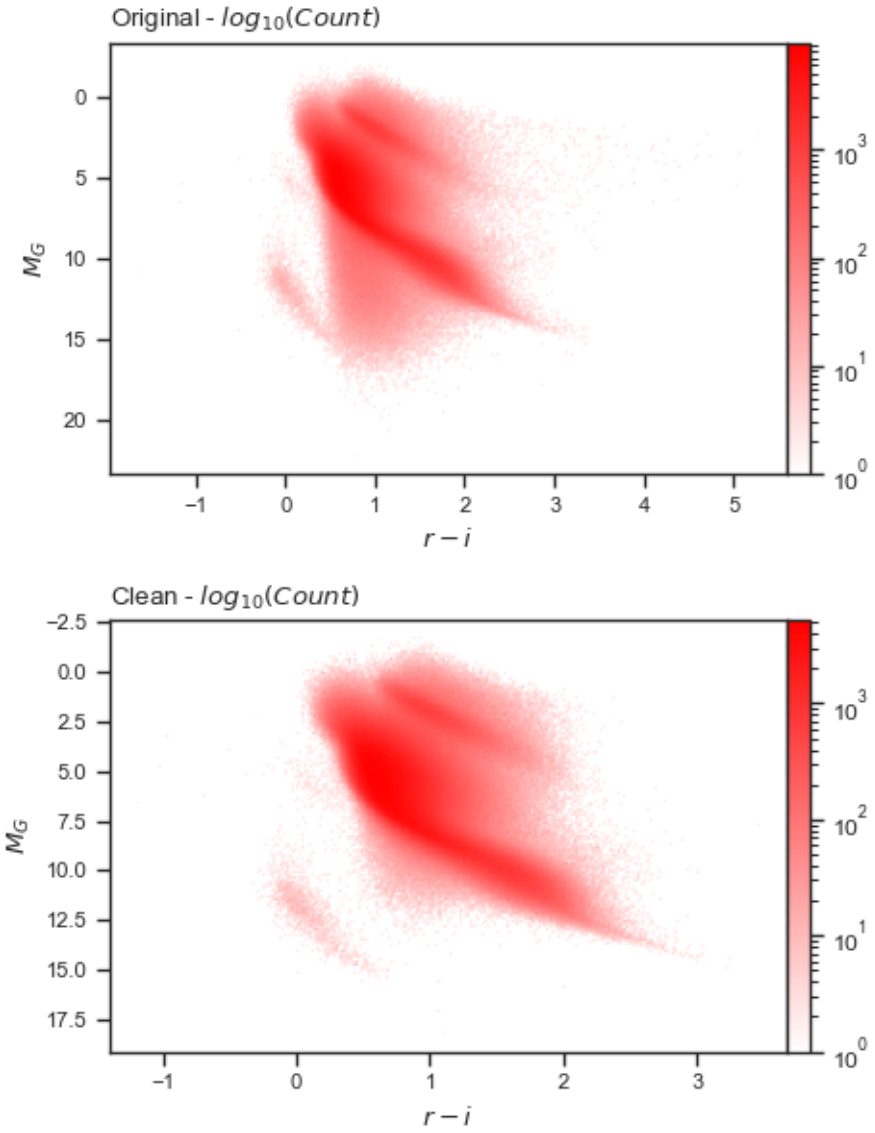


Figure 3.1: The upper plot shows the original data in a colour-magnitude log density plot, whilst the lower plot shows the data with the cuts applied. Noteworthy are the removed sources sitting between the main-sequence and the white dwarf population. There is no known population that sits in this region and these were most likely due to the spurious parallax measurements in Gaia. Removing these gives a much cleaner colour-magnitude plot.

IPHAS were used along with the absolute magnitude, M_G, from Gaia. Figure 3.2 shows the sources on a colour-magnitude and colour-colour plot.

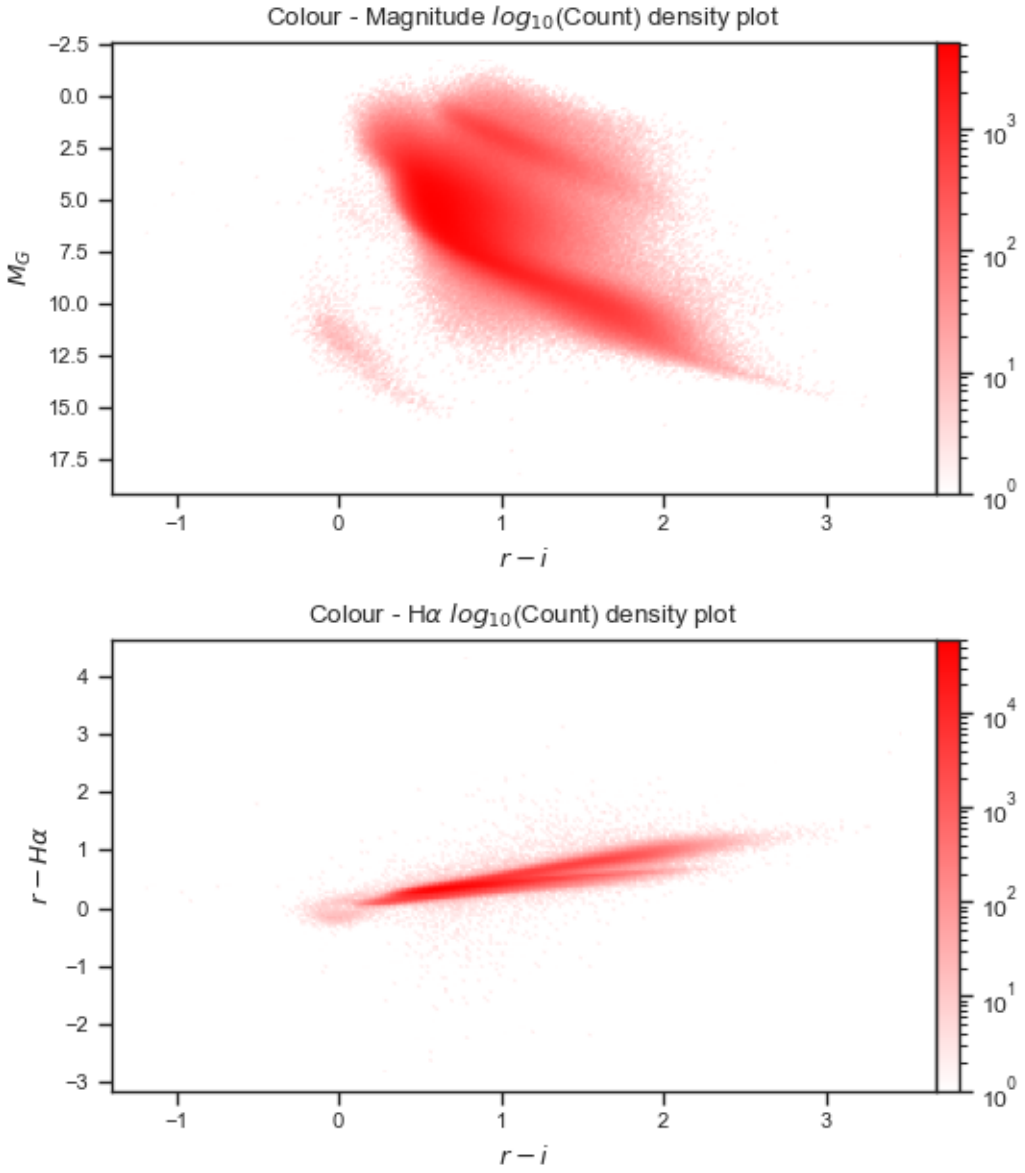


Figure 3.2: Log density plot of the cleaned data. The upper plot shows the colour-magnitude plane of the three dimensional space used. The lower is a colour-colour plot, with the upper branch containing the unreddened main-sequence, the lower branch contains the giants and reddened main-sequence; with the white dwarf population visible as a low density blob on the left side where both branches converge.

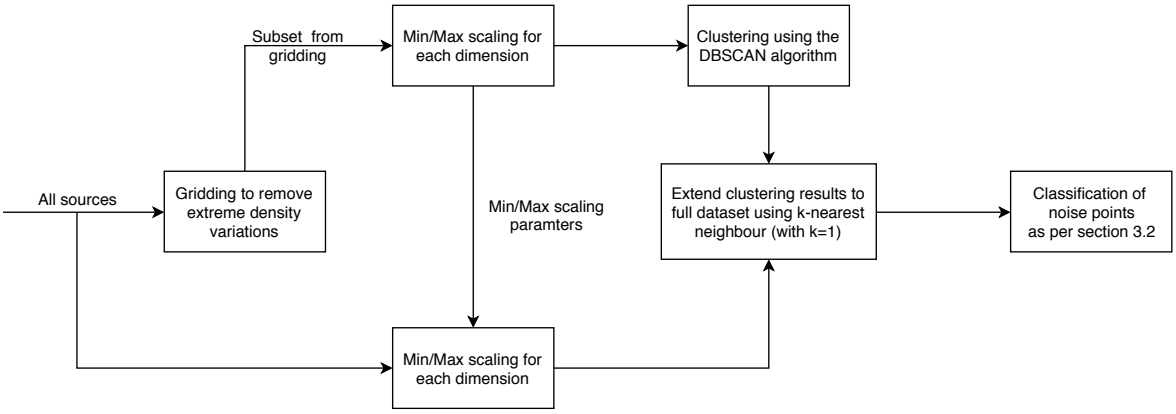


Figure 3.3: Visualisation of the main steps of the selection algorithm

3.2 Overview of Selection Algorithm

The diagram in figure 3.3 visualises the general steps of the selection algorithm, with the sections below providing more details for each step. Once the DBSCAN clustering, has been run and the labels (i.e. which group a particular source belongs to) have been applied to the full dataset, the actual selection of the emitters is performed, which is covered in detail in section 3.7.

3.3 Gridding

As figure 3.2 shows, the density variations in the three dimensional space are significant. There are two main clusters: one consisting of the pre-main sequence, the main sequence and giants; and the other consisting of the white dwarf population.

As explained in section 2.6.2, DBSCAN does not work well with clusters of varying density; therefore the density variations are flattened. Reducing the number of sources from $\sim$7.3 million to 267,237 also makes working with the data much easier (load times, manipulation, etc.).

Removing the large density variations was done by splitting the 3D space into cells and checking the number of data points in each cell against a threshold value; if below the threshold, all points in the cell are kept, otherwise the threshold number of points are selected randomly from the cell. The two parameters required for this step are: threshold value for the number of sources; and the number of cells along each dimension.

It is important to note that with a high enough resolution (i.e. number of cells per dimension), no "information" is lost when identifying $H\alpha$ emitters; only regions of the largest density (such as the main sequence and giants branches) are affected. As long as the density variations between high density regions and low density regions are maintained, and only a reduction in

density the variation between the regions is achieved, no relevant "information" is lost.
Determining the parameters for this step was done by evaluating a large number of parameter combinations and comparing and evaluating these using plots such as shown in figure 3.4. These were evaluated by looking for a similar density between the main sequence/giants region and the white dwarf population, while still maintaining the high and low density boundaries.
This reduction in number of sources allowed for fast and easy running of DBSCAN on a normal desktop machine, allowing evaluation of many DBSCAN parameter combinations. The full details of the DBSCAN parameter selection is covered in section 3.5. For gridding, the parameters used were 150 cells per dimension for each of the three dimensions, with a threshold of ten sources per cell. This parameter combination gives an upper limit of $150^3 \times 10 = 33.75 \times 10^6$ data points, given a uniform dataset with a density exceeding the threshold per cell. For this dataset, the total number of sources selected with these parameters is 267,237.

3.4 Min/Max Scaling

As DBSCAN defines the neighbourhood using a distance parameter/threshold ϵ, it is important that all dimensions span the same range. Otherwise the neighbourhood in a 2D case is no longer a circle but instead becomes elliptical if the dimensions span different data ranges; the same applies for higher dimensional spaces. In order to prevent this, the data is min/max scaled using the MinMaxScaler from the scikit-learn library [54]. The min/max scaling equation to give a range of (0, 1) is

$$X_{scaled} = (X - X_{min})/(X_{max} - X_{min}) \tag{3.1}$$

As shown in figure 3.3 the min/max scaling is initially only done on the subset that is used for DBSCAN, and, before the labels are applied to the full dataset via k-nearest neighbour, the full dataset is min/max scaled with the parameters (X_{max} and X_{min}) from the subset. Note: All figures, except of figure 5.9, from here on, are in terms of the scaled dimensions instead of their unscaled values.

3.5 DBSCAN

The density-based clustering algorithm DBSCAN, described in detail in section 2.6.2, identifies clusters that meet a density threshold defined by the two parameters $MinPts$ and ϵ. All data points that are not identified as part of a cluster are labelled as noise points. The "noise" points can be thought of as outliers with respect to their local neighbourhood. Given that the aim is

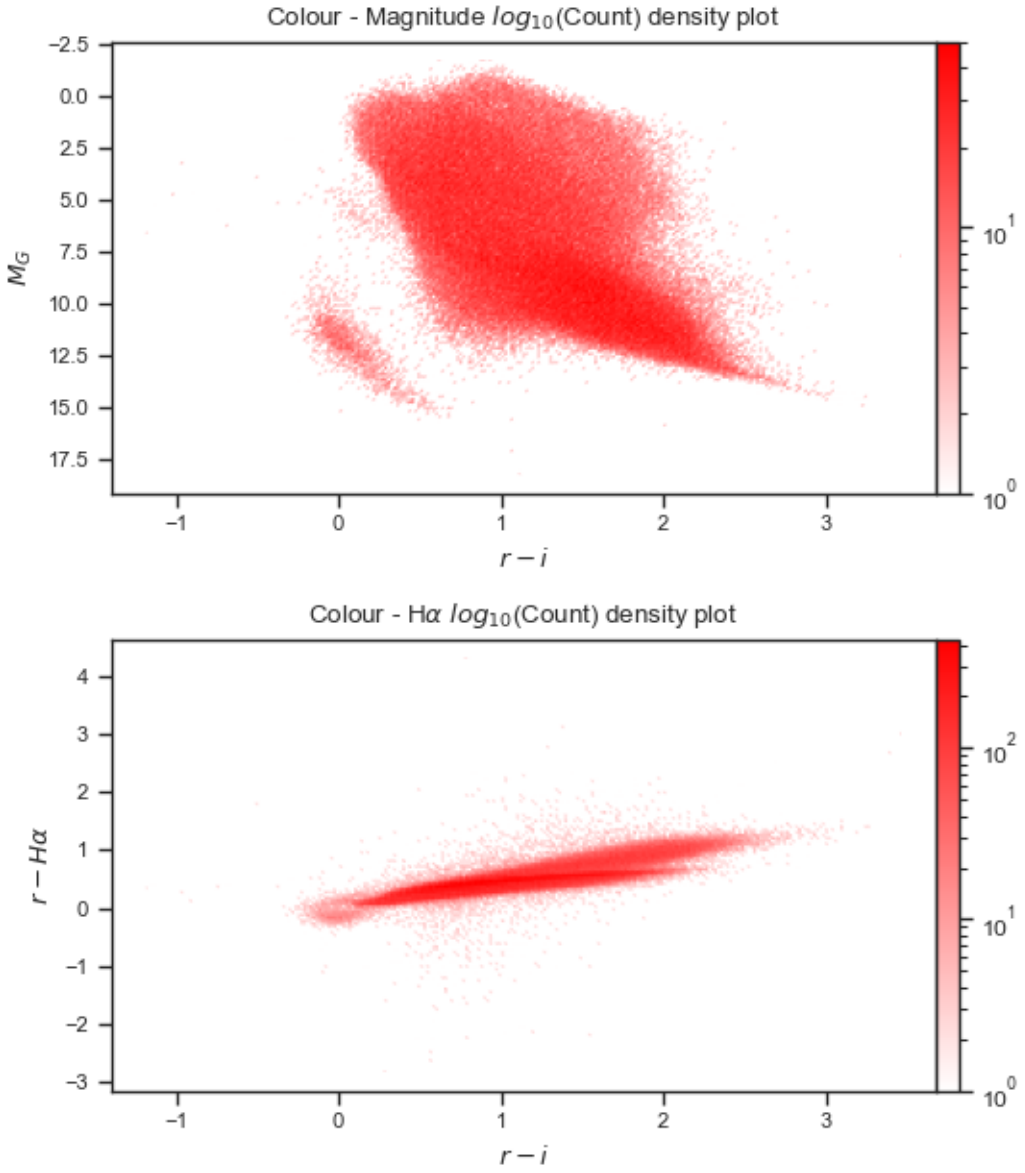


Figure 3.4: Results of gridding using 150 cells in each dimension with a threshold of ten sources per cell.

to identify excess $H\alpha$ sources, which can also be thought of as outliers in the $H\alpha$ dimension with respect to their local neighbourhood (in the CMD plane), the DBSCAN is used to find all local and global outliers, which are then the base set for selecting $H\alpha$ emitters. This reduces the number of potential excess $H\alpha$ sources dramatically, which in turns allows running the selection algorithm described in section 3.7 on all of these sources.

The DBSCAN algorithm requires a way to find sources that are close to a given point to identify its neighbourhood. This can either be done by using a distance matrix or a structure, such as k-d tree [8]. A distance matrix for n number of points and d dimensions has n^d number of entries, in other words memory requirements are $\mathcal{O}(n^d)$, compared to a k-d tree's memory usage,

which scales as $\mathcal{O}(n)$. Therefore it is advisable to use structures such as a k-d tree. However, the DBSCAN implementation by scikit-learn [53] only supports a distance matrix stored in memory, which is unsuitable even for the reduced number of data points as it requires

$$\frac{(n^d) \times 32}{10^6} MB$$

of memory when using single floating point precision (32 bytes). One of the other limitation of DBSCAN is that the original algorithm is limited to a single process and does not scale to multi-processing, due to sequential access limitations. This means that the processing of a large datasets takes a significant amount of time and having a multi-core computer to run it on does not help. Running DBSCAN on the full dataset was attempted. However, after 2 days of running the process was stopped. Modifications of DBSCAN exist that allow parallelisation, such as PDSDBSCAN [42] and [3]. However, reducing the dataset via gridding and running the original DBSCAN algorithm on a standard Linux desktop using the implementation from the PyClustering library [38], which uses a k-d tree by default, was sufficient and simple and therefore used. PyClustering is a python library, but it also allows running of the DBSCAN algorithm using a C++ core, which offers significant performance improvements. Running on a standard ubuntu desktop (Intel Xeon CPU E3-1240 v3 @ 3.40 GHz × 8, 16 GB of memory) on a reduced dataset of 267,237 data points, the run time was less than 5 minutes using the PyClustering library.

The behaviour of the DBSCAN algorithm strongly depends on the value of the two parameters, hence finding suitable values for these is important. The original DBSCAN paper [19] specifies a method of identifying these by looking for the first valley in a *sorted k-dist graph*, a sorted graph of the distance of each point to its k-th neighbour, but in this case there was no clear first valley; this is most likely due to the large number of sources. Therefore a large number of parameter combinations were evaluated by-eye using plots such as shown in figure 3.5. It is also worth pointing out that there is no absolute correct parameter combination, as changing the parameters merely changes how aggressive or conservative the clustering is, and the correct parameters might vary depending on whether completeness or accurate selection is more important.

The main problem in finding a good parameter combination was the differences in density between the main sequence and giants cluster, and the white dwarf population. While the gridding, discussed in section 3.3, reduced this density difference significantly, it was unable to eliminate it entirely. Hence when trying to select a conservative parameter combination for the main sequence cluster, most of the white dwarf population would be labelled as "noise". One approach to solve this problem is to flatten the density surface more aggressively. However, this

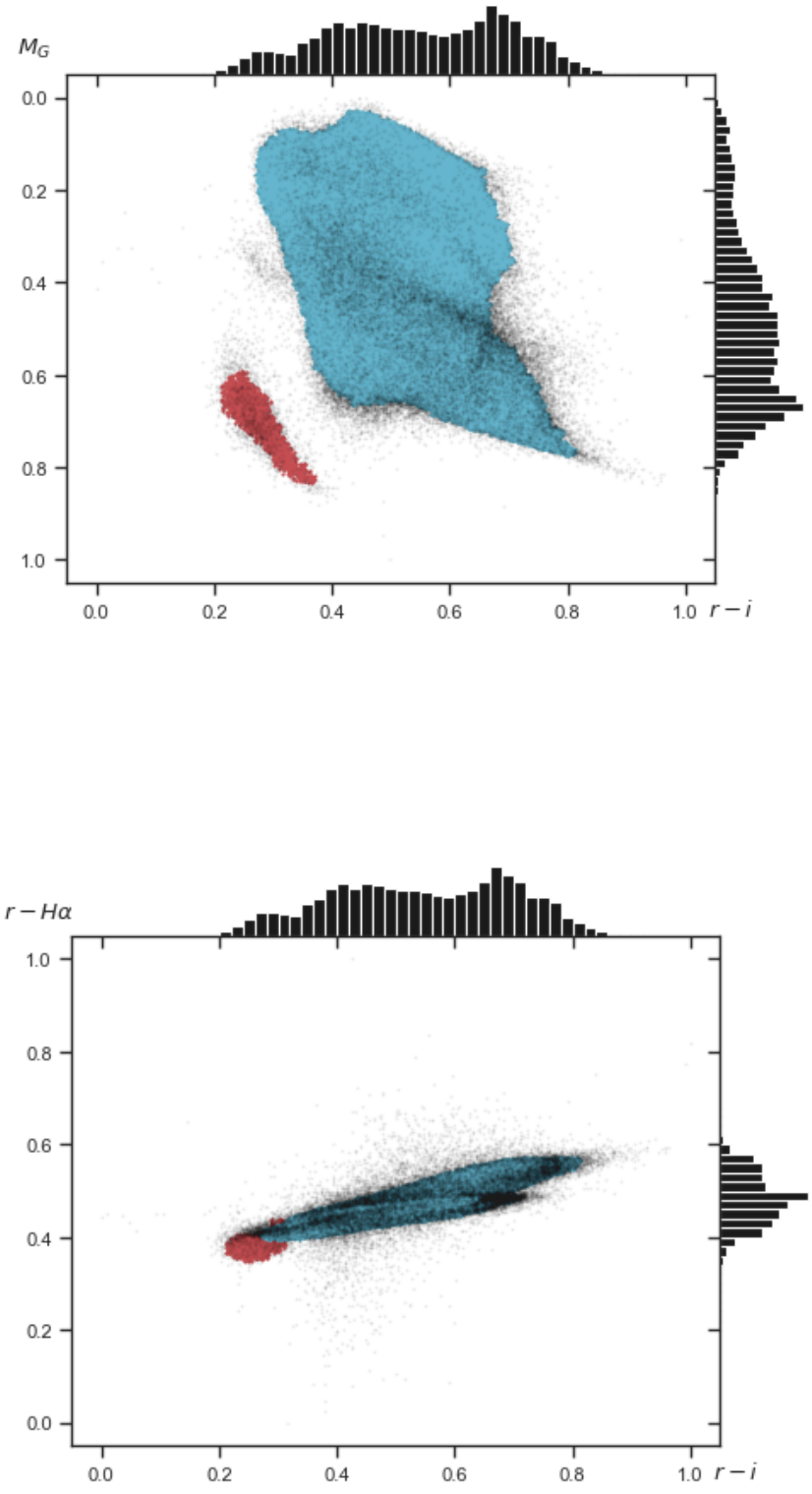


Figure 3.5: The combined results from running DBSCAN with the two different parameter combinations ($MinPts_{WD} = 20$, $\epsilon_{WD} = 0.011$ and $MinPts_{MS} = 20$, $\epsilon = 0.007$) to allow for the density variation between the white dwarf and main sequence population.

reduces the number of data points drastically, resulting in a less well-defined main sequence and giants cluster. Another problem is that some potential points of interest might be removed by this aggressive gridding. To avoid this DBSCAN is run twice with different parameters, once for identifying the main sequence and giant cluster, and once for the white dwarf cluster. The resulting two different sets of labels are trivial to combine as the clusters do not overlap. This results in a well-defined main sequence cluster without marking most of the white dwarf population as "noise". The resulting clusters and "noise" sources, for the used parameters ($MinPts_{WD} = 20$, $\epsilon_{WD} = 0.011$ and $MinPts_{MS} = 20$, $\epsilon = 0.007$), are shown in figure 3.5. This gives a main sequence and giant cluster of size $243,517$, a white dwarf cluster of size 2342, and $21,378$ "noise" sources.

3.6 Nearest Neighbour

Applying the clustering results from DBSCAN to the full dataset was done using the the k-nearest neighbour (kNN) algorithm; a brief explanation of the algorithm is given in section 2.6.1. With $k = 1$, the algorithm classifies the sources from the full dataset by assigning the label of the closest neighbouring source from the reduced dataset, with the label being the cluster membership or "noise" point classification. It is worth pointing out that kNN in this case is not used for instance based supervised machine learning; it is purely used to apply the DBSCAN results to the full dataset.

A k-value of 1 was used as larger values would mean considering more neighbours when classifying a point. This could lead to data points being incorrectly classified as cluster points at the boundary of the clusters due to the much higher number of cluster sources compared to "noise" sources. It also worth noting that applying the DBSCAN clustering results to the full dataset only results in a very minor change in the number of "noise" sources, from $21,378$ to $21,381$. The number of sources in the white dwarf cluster, changes from $2,342$ to $2,346$. This shows that reducing the number of sources, as explained in section 3.3, almost purely removed sources from the main sequence and giants cluster; explaining the small changes in the number of "noise" and white dwarf cluster points.

The full dataset was also min/max scaled before running kNN, as otherwise the sources would have been in a different data range, producing incorrect results. To ensure that the full dataset was scaled in the same way as the reduced dataset, the same instance of the min/max scaler was used (which saves the parameters X_{min} and X_{max}).

As the aim of this step is to apply the clustering from the reduced dataset to the full dataset, it

is important to check that the transferred clustering does not overextend itself. In other words, the result should be very similar to running the DBSCAN on the full dataset. To verify this, the nearest neighbour sorted distance graphs for both clusters (for the full dataset) were produced, as shown in figure 3.6, and were compared to their respective DBSCAN ϵ parameter.

The graphs show that none of the cluster points exceed their respective cluster's ϵ param-

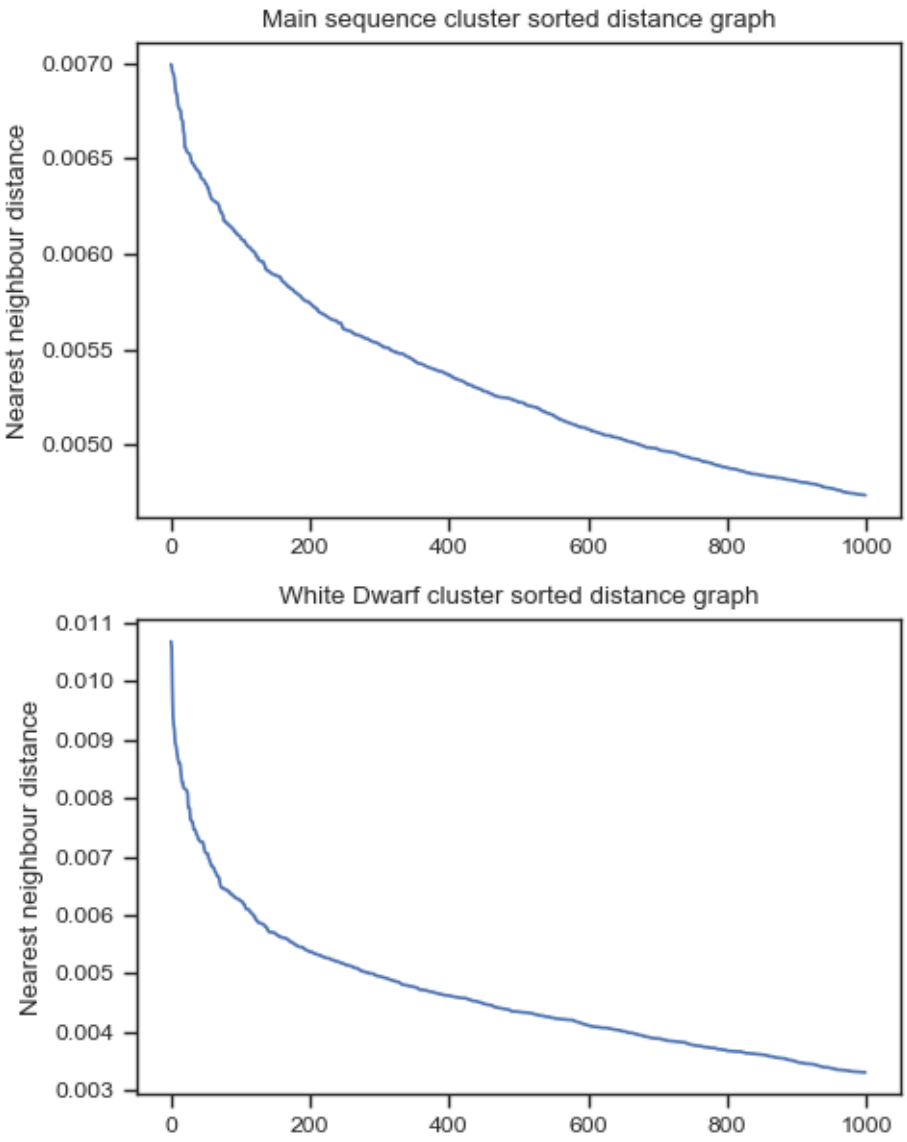


Figure 3.6: Sorted nearest neighbour distance graph for the main-sequence/giant cluster (*top*) and the white dwarf cluster (*bottom*). Each plot shows nearest neighbour distance for the first 1000 cluster sources, when sorted by the nearest neighbour distance. For both clusters the maximum nearest neighbour distance is below their respective DBSCAN ϵ parameter ($\epsilon = 0.007$ for the main-sequence/giant cluster and $\epsilon = 0.011$ for the white dwarf cluster). This indicates that applying the DBSCAN clusters to the full datasets using nearest neighbour did not result in the clusters overextending themselves.

eter. However, this does not mean that this is equivalent of running DBSCAN on the full

dataset, as this does not take into account the differentiation of core points and border points in a cluster. The results however, are expected to be very similar to running DBSCAN on the full dataset and given that the DBSCAN parameters were chosen to be conservative, i.e. lots of "noise" sources, the risk of missing a source of interest (with respect to excess $H\alpha$) is minimal.

3.7 Selection Process

Selection of the $H\alpha$ emitters from the set of "noise" sources is done with respect to a "selection neighbourhood", which is the "local neighbourhood" of a given "noise" point modified to include a proportion of the closest locus on the colour-magnitude plot. This section covers the two techniques used to choose the selection neighbourhood for a given "noise" point, which is subsequently referred to as a source of interest (SoI).

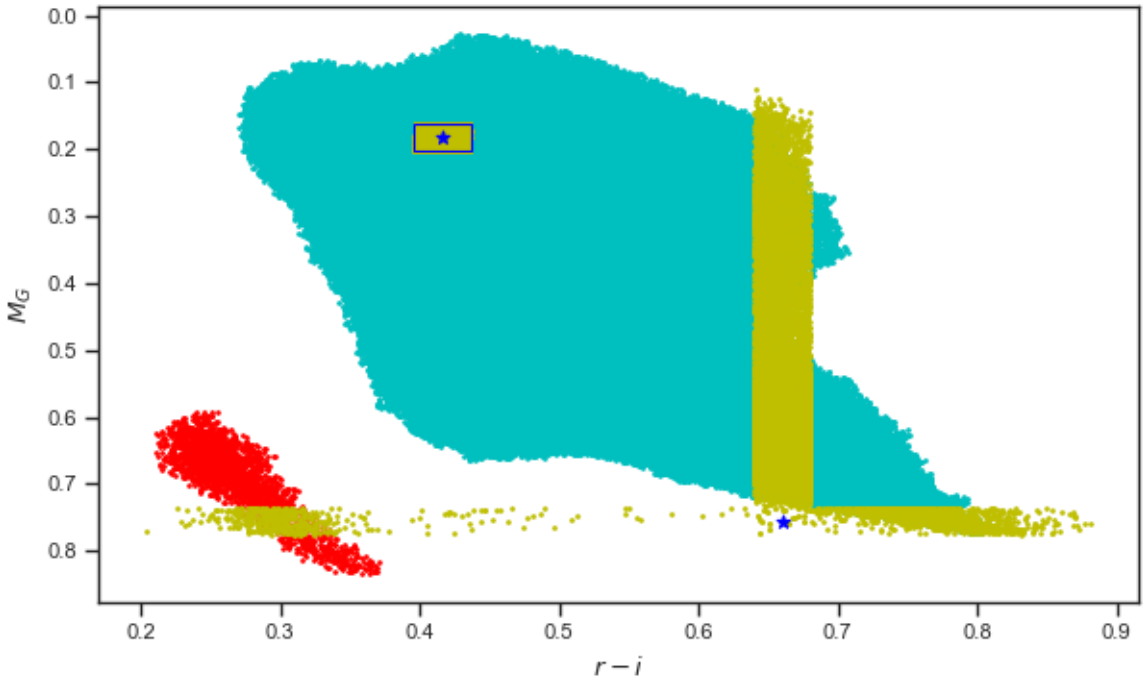


Figure 3.7: Shows the two different techniques used to determine the selection neighbourhood for a SoI. For the SoI in the upper left, highlighted by the blue star, it is found using tunnelling, as covered in section 3.7.1. For the SoI in the lower left, slicing in both dimensions of the CMD is used to find it; slicing is covered in section 3.7.2. Note: This colour-magnitude diagram does not show any "noise" sources, apart from the ones in a slice.

3.7.1 Tunnelling

Tunnelling for a SoI is done by selecting all sources that are within a rectangular area of the SoI in the CMD plane; in other words if the SoI has position $(SoI_{r\text{-}i}, SoI_{M_G})$ then all sources that are within the rectangle defined by $(SoI_{r\text{-}i} - w) < r\text{-}i < (SoI_{r\text{-}i} + w)$ and $(SoI_{M_G} - h) < M_G < (SoI_{M_G} + h)$ make up the selection neighbourhood, where h and w are the height and width of the rectangle in the scaled dimensions. It is important to note that the $r\text{-}H\alpha$ dimension is unconstrained. The height and width used was 0.04 in the scaled dimensions. An example is shown in figure 3.7, with the SoI highlighted as a blue star in the top left of the CMD. The selected sources, or selection neighbourhood, are shown in yellow, with the constraints highlighted by the blue rectangle. The associated histogram and $r\text{-}i$ vs. $r\text{-}H\alpha$ plots are shown in figure 3.8, from which it is easy to see that this SoI is most likely an $H\alpha$ emitter. The automatic selection of $H\alpha$ emitters based on the selection neighbourhoods is covered in chapter 4.

In order to perform a statistically meaningful selection of $H\alpha$ emitters using their respective selection neighbourhoods, these have to contain a large enough number of sources. However, for sources that are not above/below or close to a dense region in the CMD plane, this method will not result in a meaningful selection neighbourhood. So while tunnelling was run for every SoI, it was only used to get the selection neighbourhood for SoIs that are considered "*on cluster*". For a source to be considered "*on cluster*" the following conditions had to be meet:

- Number of sources in tunnel >= 200
- Number of cluster sources >= 100

The second condition was added to ensure that the selection neighbourhood contains a core part of the closest locus in the CMD plane, as the boundary regions of the main-sequence/giant cluster can contain enough sources for some SoIs to reach the required tunnel count without containing any cluster sources.

The single parameter of this method, the width and height of the rectangle, determines the neighbourhood in the CMD plane considered when classifying a SoI. The width/height parameter used was 0.04/0.04 as this provided a good neighbourhood selection while also allowing the inclusions of boundary sources that meet the cluster count condition.

3.7.2 Slicing

As mentioned in 3.7.1 some SoI will not meet the required conditions to get a selection neighbourhood from tunnelling; these are SoIs that sit in low density regions or at the edges of a

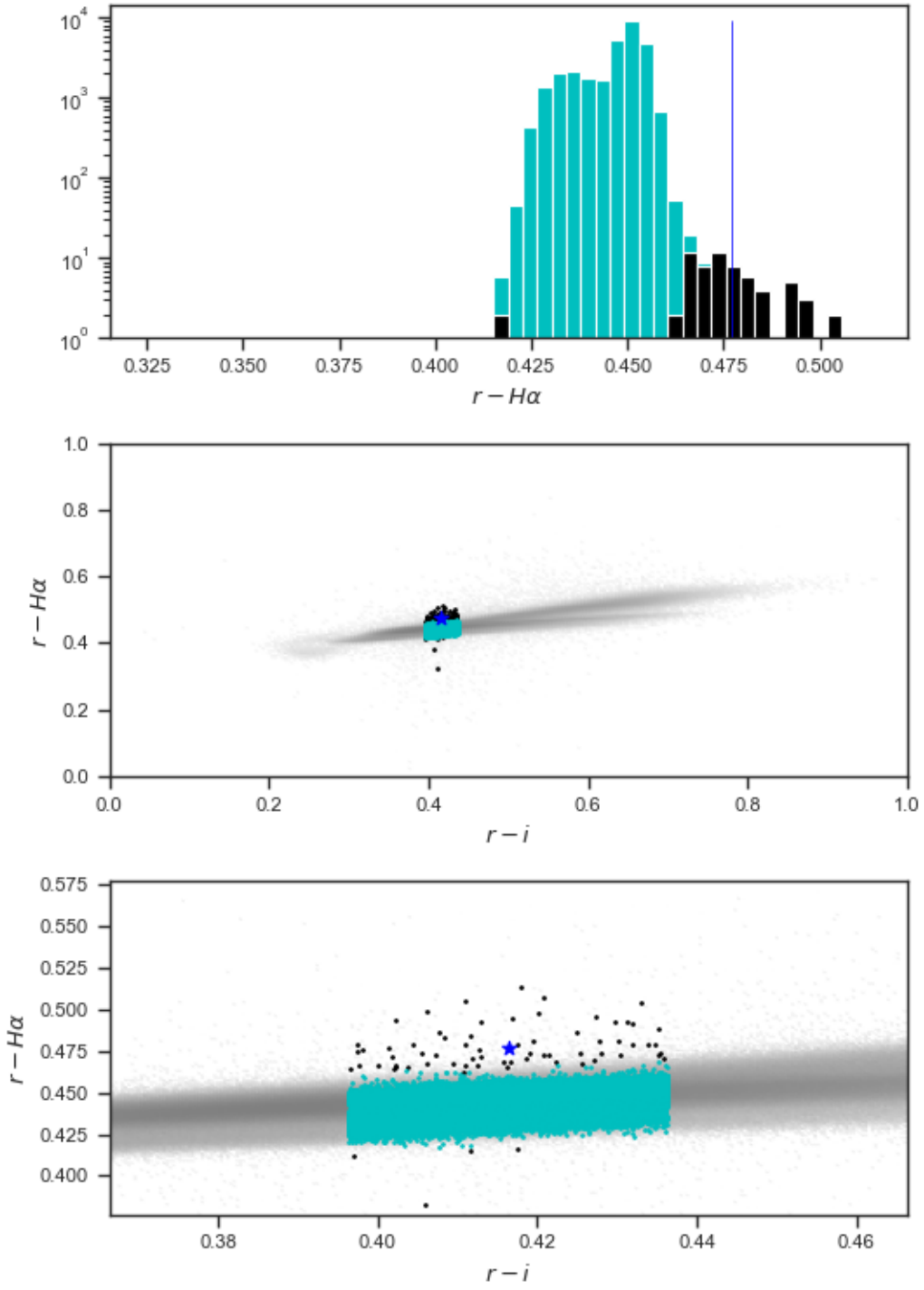


Figure 3.8: Selection neighbourhood plots for the SoI in the top left of the CMD in figure 3.7. (*Top*) The selection neighbourhood's *r*-$H\alpha$ distribution on a log scale as this SoI is in a very dense region of the CMD. (*Centre*): colour-colour plot with the cluster sources in the selection neighbourhood shown in cyan and the "noise" sources in black. All other sources are shown as a density plot to show the loci. (*Bottom*): Same plot as in the *middle* but zoomed in on the SoI, showing, more clearly how it sits above the locus in the *r*-$H\alpha$ dimension.

locus in the CMD. For these SoIs slicing is used.

Slicing is done in the CMD plane (i.e. the r-$H\alpha$ dimension is unconstrained) by selecting all sources along the slice dimension, either r-i or M_G, that meet the constrains on the other dimension. For example, slicing along the M_G dimensions for a SoI positioned at (SoI_{rmi}, SoI_{M_G}) is done by selecting all sources that meet the conditions $(SoI_{rmi} - \textit{slice width}) < x_{rmi} < (SoI_{rmi} + \textit{slice width})$. An example of slicing along the dimensions $r - i$ and M_G is shown in figure 3.7, where the SoI is shown as a blue star in the bottom right of the CMD and the selected sources are in yellow. The top plots in figure 3.9 show the scatter plots for the slices, with the slice dimension on the x-axis and r-$H\alpha$ on the y-axis. These plots give a good indication of whether a SoI is an emitter or not. However, in order to select an $H\alpha$ emitter automatically, the selection neighbourhood has to be determined. If one were to determine if a SoI is an emitter based on the top two plots in figure 3.9 and the CMD in 3.7, one would most likely find the slice that contains the closest population and then view the respective *slice dimension* vs r-$H\alpha$ plot and compare the SoI against the a portion of the closest locus. Therefore, this is how the selection neighbourhood is determined for SoIs that don't meet the tunnelling conditions. The important point to note is that the selection neighbourhood needs to include a core part of the closest locus.

The first step to determining the selection neighbourhood is to choose the correct slice to use, which is done using the following steps:

1. For each slice:
   (a) Check that number of points in the slice > 200;
   (b) Calculate the distance of each point with respect to the SoI;
   (c) Find the closest cluster, for which there are at least 100 cluster points in the slice; and
   (d) Return the label and distance of the closest cluster (1c) point.
2. Select the slice with the closest valid cluster point.

With the correct slice selected, the next step is to identify the "local" neighbourhood that encompasses a large enough core portion of the closest locus (i.e. cluster) from the selected slice. This selection of points has to meet the following conditions:

- Number of data points in the selection > 200; and
- Number of cluster points (from the cluster in 1c above) > 100.

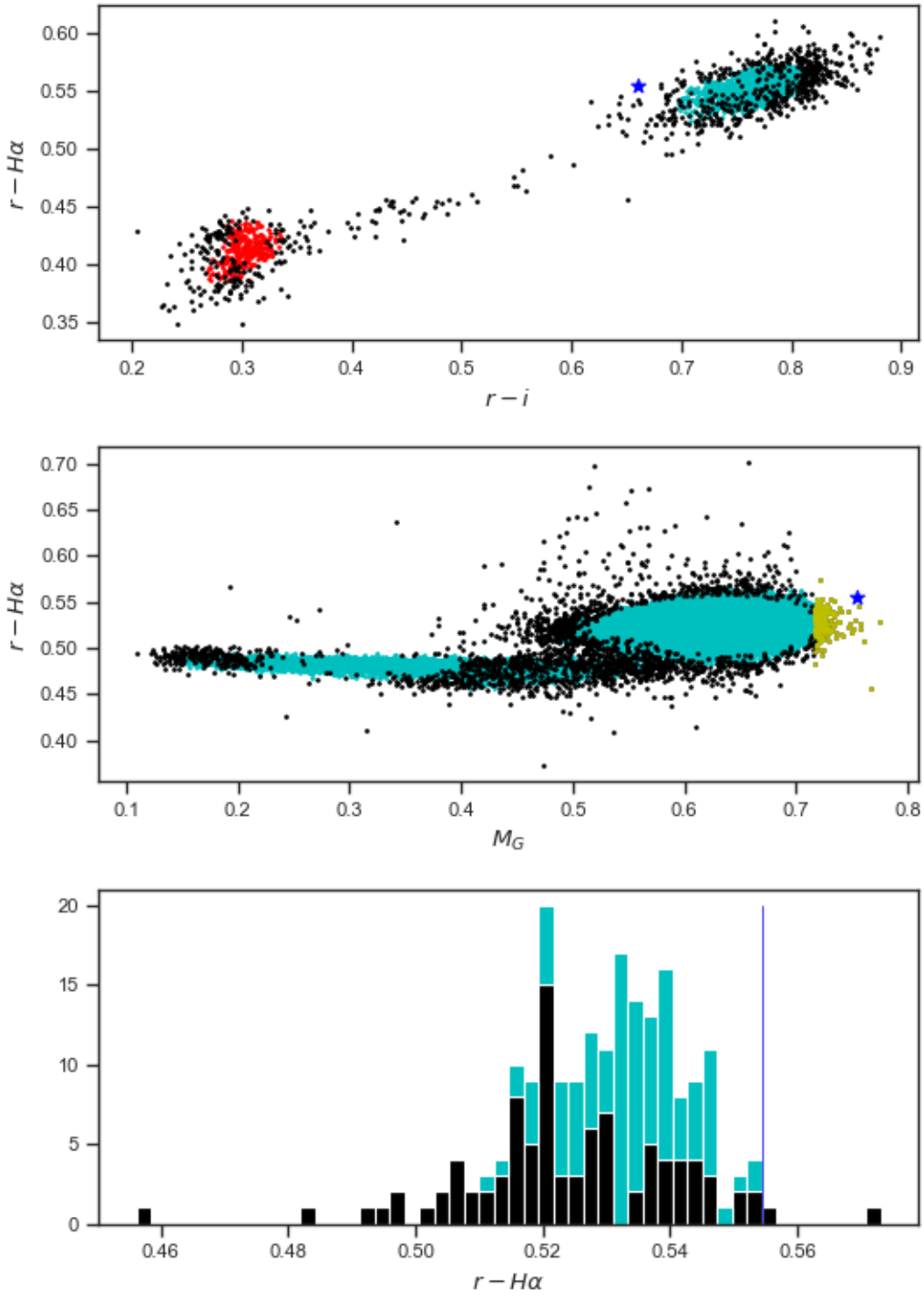


Figure 3.9: Slicing plots for the SoI in the bottom right of the CMD on figure 3.7. (*Top*) Scatter colour-colour plot of the sources in the *r-i* slice. In this case the relevant slice is M_G as it intersects with the closest locus on the CMD. (*Centre*) Colour-colour scatter plot of the M_G slice, with the selection neighbourhood highlighted in yellow. (*Bottom*) Histogram showing the *r-H*α distribution of the selection neighbourhood in the *centre* plot. From this, and the *centre* plot, it is apparent that this source is most likely not an $H\alpha$ emission line source.

For the example in figure 3.9 the selection neighbourhood points are shown in yellow in the middle plot. With these selected points it is then possible to score (covered in section 3.7.4) this SoI the same way as tunnel points.

3.7.3 Manual Reference Point

For a very small subset of SoIs (∼200) neither slicing nor tunnelling provides a sufficient selection neighbourhood, as these SoIs are not in a dense region of the CMD and the slices do not intersect with their respective closest clusters. These SoIs are shown in figure 3.10 and found using the selection areas shown by the blue lines. For each of the three groups a reference source (magenta stars) was determined, and the tunnel sources from the three reference sources were then used as the selection neighbourhoods for the different areas. The tunnelling area was slightly increased for two reference points, as they are in low density regions. With the selection neighbourhoods determined, these points can be scored and $H\alpha$ emitters selected the same way as for all other SoIs.

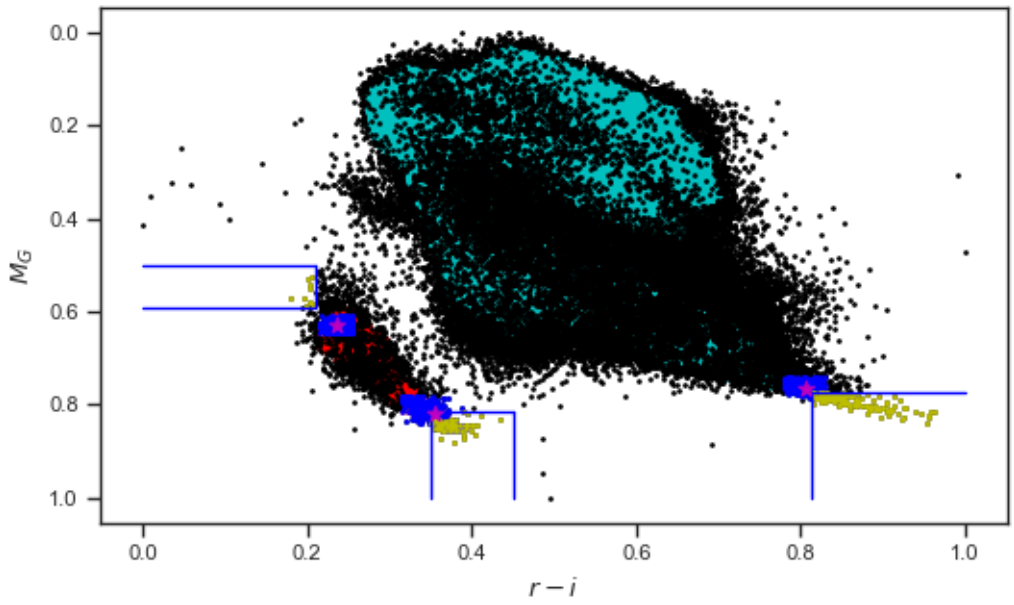


Figure 3.10: CMD plot with the SoIs that require manual reference points shown in yellow and selection criteria shown as blue lines. The selection neighbourhood sources for these SoIs are shown in blue, with the reference point shown as a magenta star. These SoIs require a manually selected reference point as they do not meet the tunnelling count conditions and their slices do not intersect with the closest cluster, which could therefore lead to incorrect selection of $H\alpha$ emitters.

3.7.4 Scoring

Qualitatively it is reasonably easy to see if an SoI is an outlier in the r-$H\alpha$ dimension with respect to the selection neighbourhood using a histogram such as the ones shown in figure 3.8 and 3.9. To reiterate from the previous sections, the selection neighbourhood for a given SoI consists of the sources against which it is scored, and is determined by either tunnelling, slicing or manual reference point depending on where it sits in the CMD, as discussed in sections 3.7.1, 3.7.2 and 3.7.3. With the selection neighbourhood it is then possible to quantitatively describe to what degree a SoI is an outlier in the r-$H\alpha$ dimension. The following sections give some details on the different methods of scoring that were considered, along with the two that were used in the end. It is important to note that all scores use the selection neighbourhood of a SoI to score it and which approach was used to determine the selection matter does not affect the score; therefore the following section is in terms of the r-$H\alpha$ dimension for the selection neighbourhood of the SoI in question.

Median/IQR

If the distribution of points in the selection neighbourhoods along the r-$H\alpha$ dimension followed a normal distribution, the first choice would be to use the mean and standard deviation to describe it and use a σ-cutoff to select emitters. However, as the plots in figure 3.11 show, the distributions do not follow a normal distribution. Therefore, for the SoI in question the median (M) was calculated along with the interquartile range (IQR), as shown in figures 3.12 and 3.13. The simple score

$$M/IQR_{SoI} = \frac{Median - SoI_{r\text{-}H\alpha}}{IQR}, \tag{3.2}$$

was then attempted to be used for selection of $H\alpha$ emitters.

However, due to large variations in the shape of the distributions, as shown in figure 3.11, this scoring method was unsuitable for selection of $H\alpha$ emitters. Figures 3.12 and 3.13 illustrate the shortcomings: both SoIs have a Median/IQR score of $6.1 - 6.2$, but the SoI in figure 3.12 is considerably less of an outlier in the r-$H\alpha$ dimension compared to the SoI in 3.13. In addition the IQR factor results in SoIs with a similar degree of outlier-ness, having vastly different scores due the IQR varying greatly. In Figure 3.14 the r-$H\alpha$ distance of the closest cluster source to the SoI is plotted against the Median/IQR for every SoI. Taking the closest cluster source distance as a proxy for r-$H\alpha$ outlier-ness, one can see that the score varies greatly for a constant distance, such as 0.05. Therefore it was concluded that the combination of median and IQR is not able to describe the different r-$H\alpha$ distributions of selection neighbourhoods in a consistent

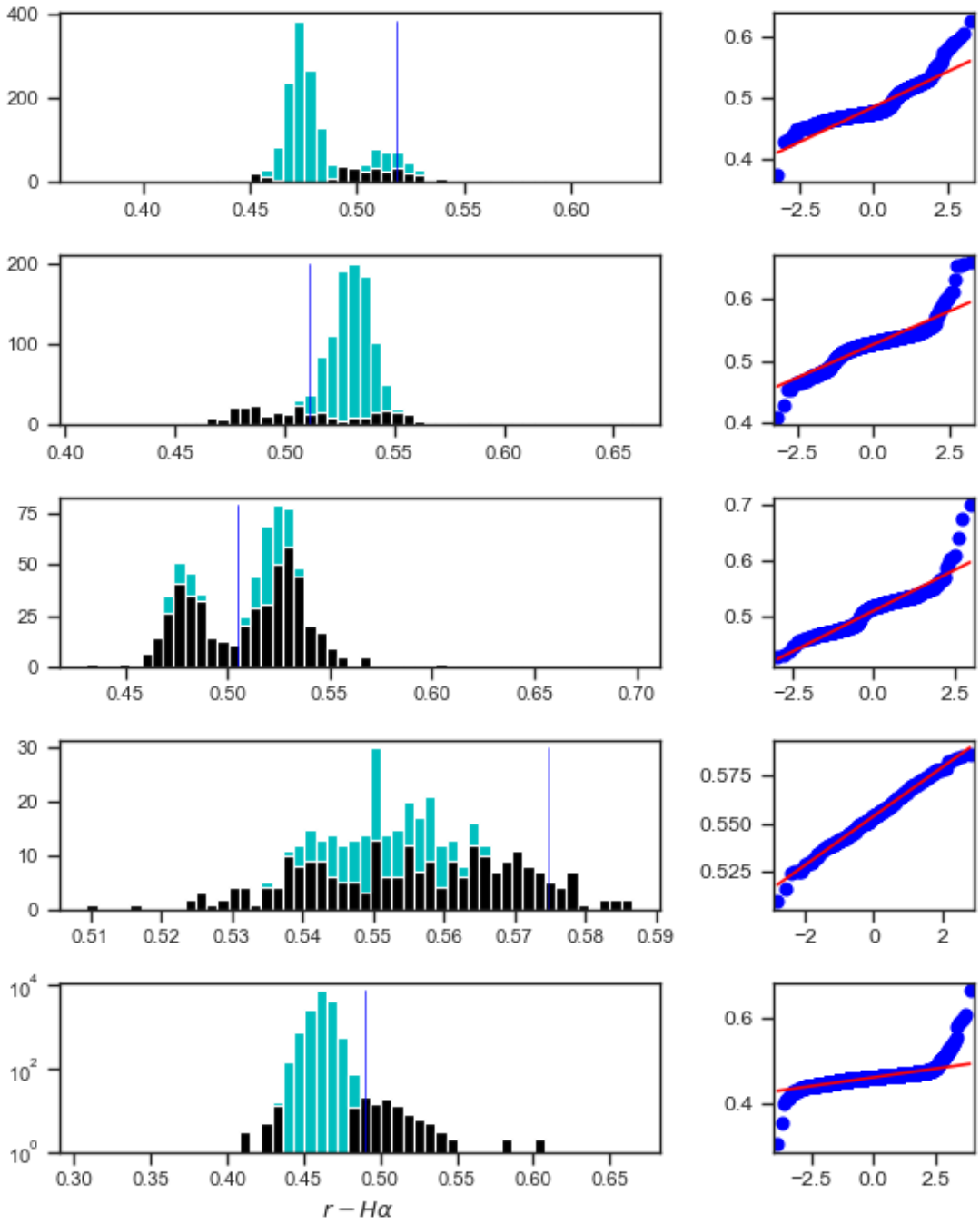


Figure 3.11: Some r-$H\alpha$ selection neighbourhood distribution plots for different SoIs. The Q-Q plots on the right are shown for comparison with a normal distribution, represented by the red line with the actual distribution shown in blue. Of note are the large differences in the distribution shapes and sizes. *Top/Center-Top* Both are SoIs situated above the lower main sequence, with the shape of the distributions varying significantly. The histogram also shows that the *top* and *Center* selection neighbourhood contains two well defined loci, whereas the *center-top* one has only a single well defined locus. *Center/Center-Bottom* Two SoI on the border of the main-sequence/giant cluster; unlike SoIs sitting directly above a cluster, in the r-$H\alpha$ dimension, like the one in the *bottom* plot, these have a much smaller number of sources in their selection neighbourhood with the "noise" sources making up a significantly larger proportion.

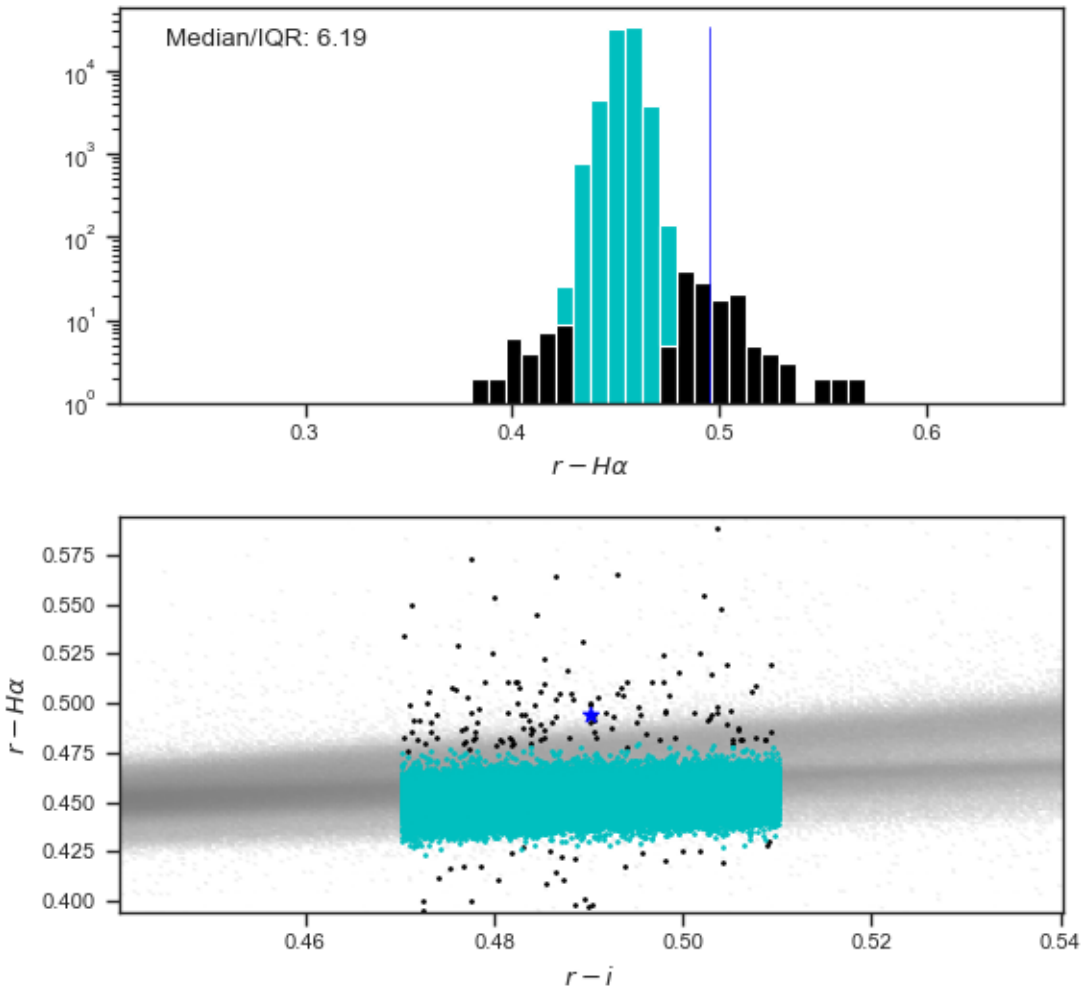


Figure 3.12: Example of a SoI which sits well above the locus from its selection neighbourhood. The median/IQR score is shown in the top left to allow comparison of two SoIs that have a very similar score but very different $H\alpha$ values with respect to their selection neighbourhood. The other SoI is shown in figure 3.13

manner, resulting in scores that are unsuitable for the selection of emitters.

Empirical probability score

For a given SoI the empirical probability (EP) score is the proportion of selection neighbourhood sources which have a smaller or equal r-$H\alpha$ value. It gives a measure of how extreme a SoI's r-$H\alpha$ value is with respect to the selection neighbourhood sources. Defined as

$$EP_{SoI} = P(X \leq SoI_{r\text{-}H\alpha}), \tag{3.3}$$

it is the value of the cumulative distribution function for the selection neighbourhood points at $SoI_{r\text{-}H\alpha}$. However, it is important to note that this is only with respect to the selection neighbourhood. Thus, if the proportion of locus points is very large, or the range of r-$H\alpha$ values

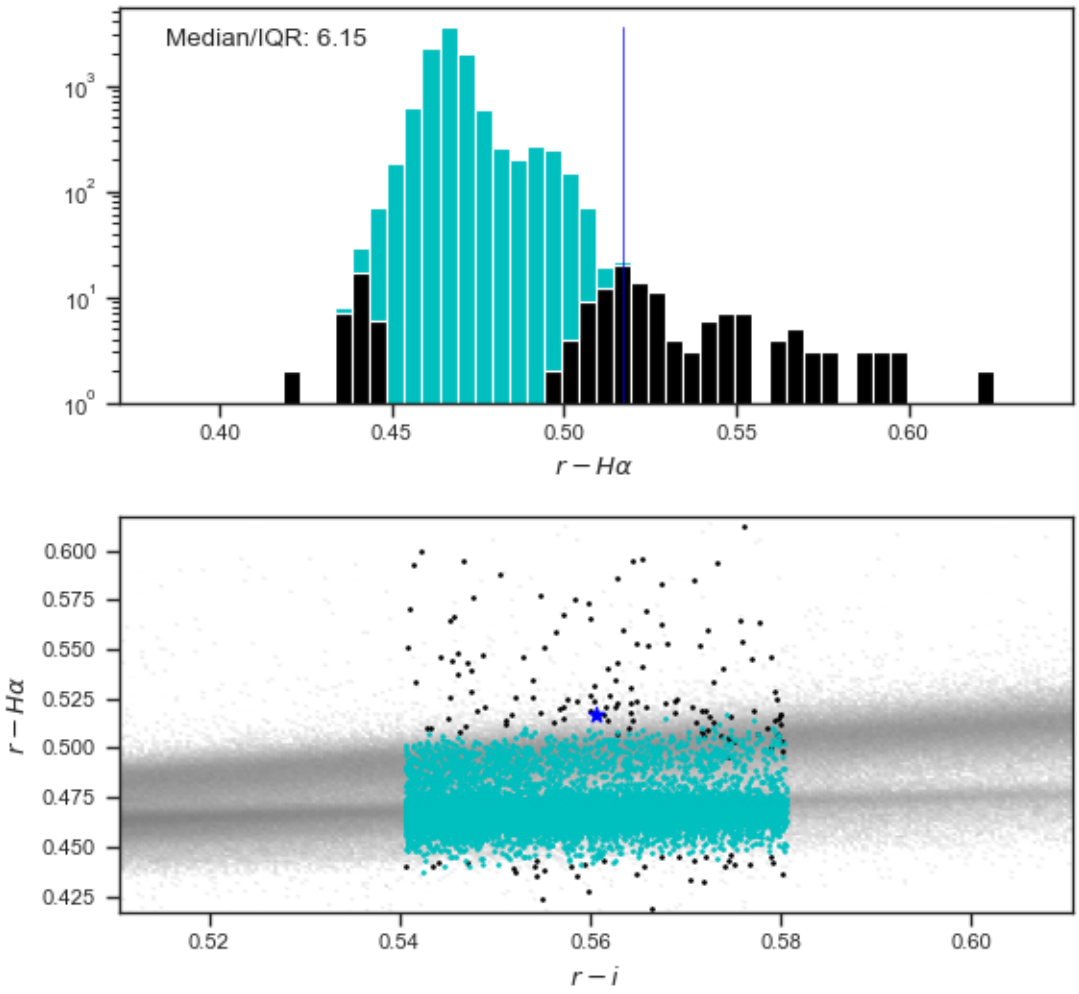


Figure 3.13: Example of SoI which sits much closer to the locus in its selection neighbourhood compared to the one shown in figure 3.12; yet its median/IQR score is almost exactly the same.

is very limited, any point that is on the tail end of the distribution (which could still be part of the locus), will have a EP value close to 1.0. This can be seen in figure 3.15. The SoI is on the outer edge of the locus and could potentially be considered as an emitter, but visually comparing it to the other sources in the selection neighbourhood, it is obvious that there are many other sources with more extreme r-$H\alpha$ values, yet the SoI has an EP score of 1.0 (rounded). On the other hand, if the number of selection neighbourhood points is low then the EP score might be low even if an SoI has an extreme r-$H\alpha$ value. This effect is due to the extreme variations in the number of sources in the selection neighbourhoods. Some of the SoIs in the dense region of the CMD have selection neighbourhoods on the order of 10^6 sources, whereas in less dense regions, or on the edges of the populations, some only have the minimum of 200. This means that this score is unsuitable for selection of $H\alpha$ emitters without addressing these large differences in selection neighbourhoods.

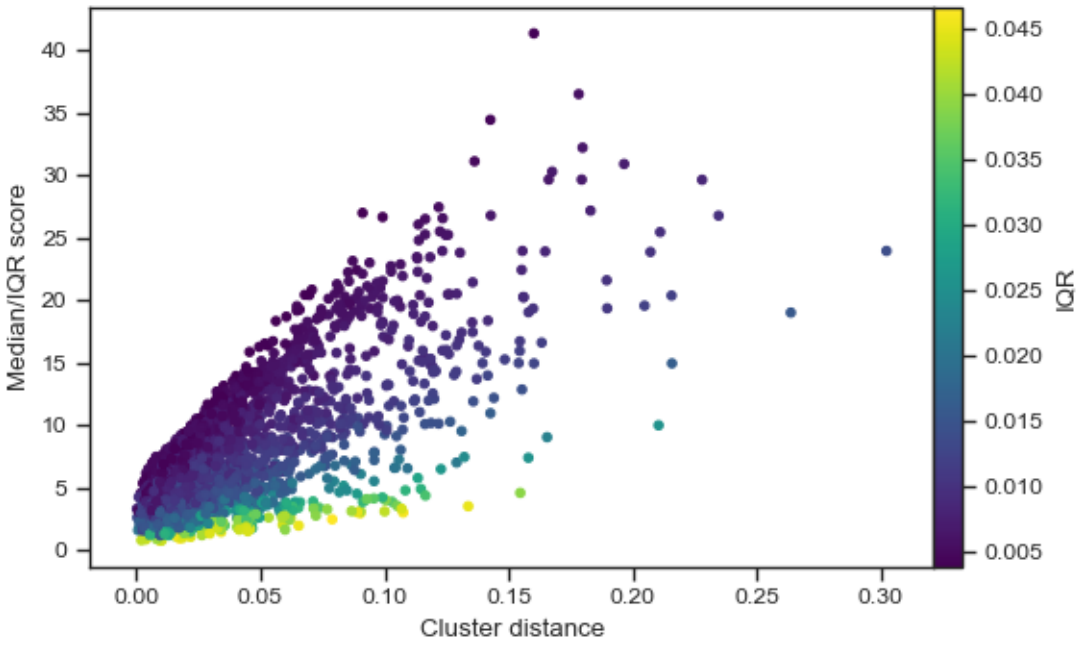


Figure 3.14: Plot of SoI's $H\alpha$ distance to the nearest cluster source in their respective selection neighbourhood versus the median/IQR score, with the colours visualising their selection neighbourhoods IQR value. Of note is the large spread of the median/IQR score at a constant cluster points distance, for example at a cluster distance of 0.05 the score varies from ∼1-15.

Weighted empirical score

The main limitation of the previously discussed EP score is the large differences in the number of sources in the selection neighbourhoods. The weighted empirical score (WEP) is a modification to the EP score that addresses this issue. An example of this problem is shown in figures 3.15 and 3.16 where there is a difference in the number of sources of four orders of magnitude. This has a large effect on the EP score. The SoI in figure 3.15 has an EP score of 1.0, whereas the SoI in figure 3.16 has a score of 0.96, even though the plots clearly show that the SoI in figure 3.16 is much more likely to be an emitter.

These large differences in the number of sources are caused by the extreme differences in density on the CMD, and as sources in dense regions are cluster points, the weighted EP score adds weights to the r-$H\alpha$ values of cluster sources if the number of cluster points in the selection neighbourhood exceeds a threshold. The weighted EP score is calculated in the same way as the EP score, except that a modified CDF is used, defined as

$$CDF(x) = \sum_{i=1}^{i=n} \begin{cases} C_{i,r\text{-}H\alpha}, \text{if “noise”} \\ w \times C_{i,r\text{-}H\alpha}, \text{if cluster,} \end{cases} \tag{3.4}$$

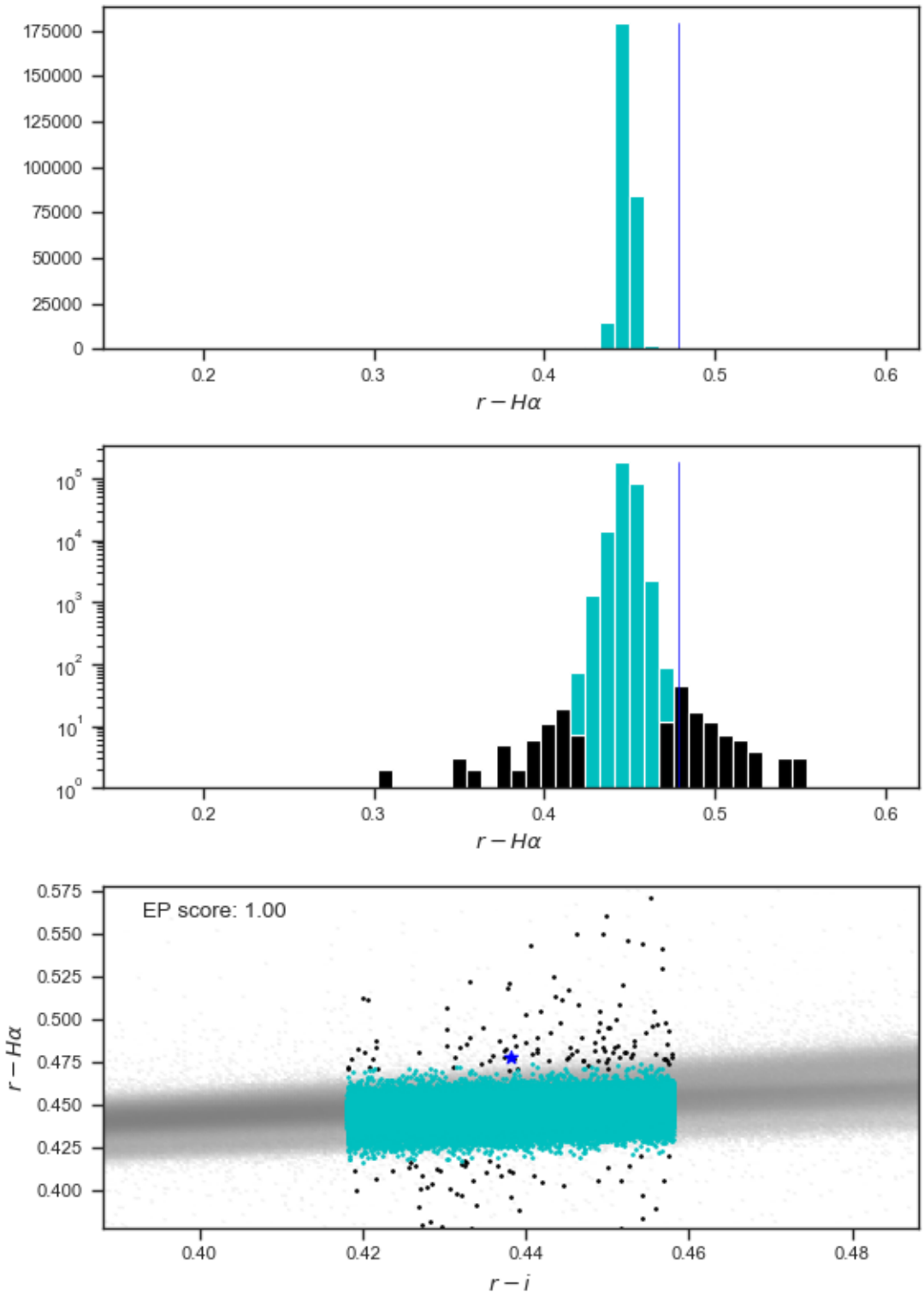


Figure 3.15: Plots for a SoI with a high density selection neighbourhood. (*Top*) The *r*-$H\alpha$ distribution of the selection neighbourhood. (*Center*) The same distribution on a log scale. (*Bottom*) Zoomed in colour-colour plot of the selection neighbourhood sources, with the SoIs associated EP score shown in the top left. Of note is the reasonably small distance between the SoI and the locus, making this a possible emitter, but its $H\alpha$ value is certainly not extreme with respect to its selection neighbourhood, yet its EP score is 1. Comparing this to the SoI in figure 3.16 shows the problem with the EP score.

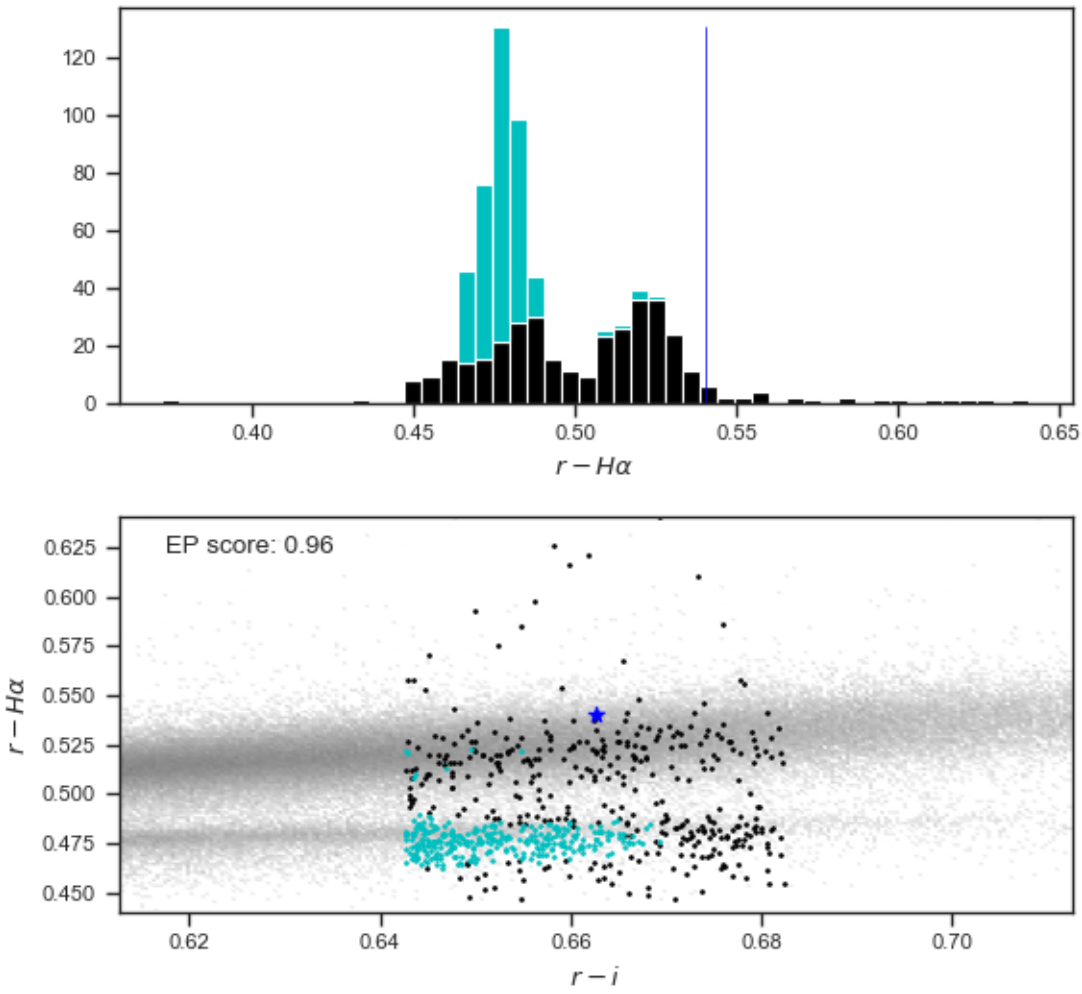


Figure 3.16: *r-Hα* distribution and colour-colour plot for a SoI in a much lower density region in the CMD. The SoI is almost certainly an $H\alpha$ emitter, yet its EP score does not reflect that accordingly, especially when compared to the SoI in figure 3.15.

where w is the weights for the cluster points defined as $w = \frac{\text{threshold}}{\text{number of cluster points (in neighbourhood)}}$, i.e. effectively reducing the effect of the cluster sources to a maximum combined total of *threshold* number of sources. The difference in the CDF due to the weights can be seen in figure 3.17. The top plot is without weights and the full probability range is completely dominated by the cluster sources, with "noise" sources at the tails making up virtually none of the probability range; hence defeating the purpose of the EP score. The main issue is not that there is a difference in the number of cluster sources, but that these difference are several orders of magnitude, thus reducing the differences to be less than an order of magnitude results in a score that is able to capture the degree of *r-Hα* deviation of SoIs with respect to their selection neighbourhood sources in a much more consistent manner. The effect of adding weights is shown in the bottom CDF plot in figure 3.17, with the cluster points only covering a portion of the probability range.

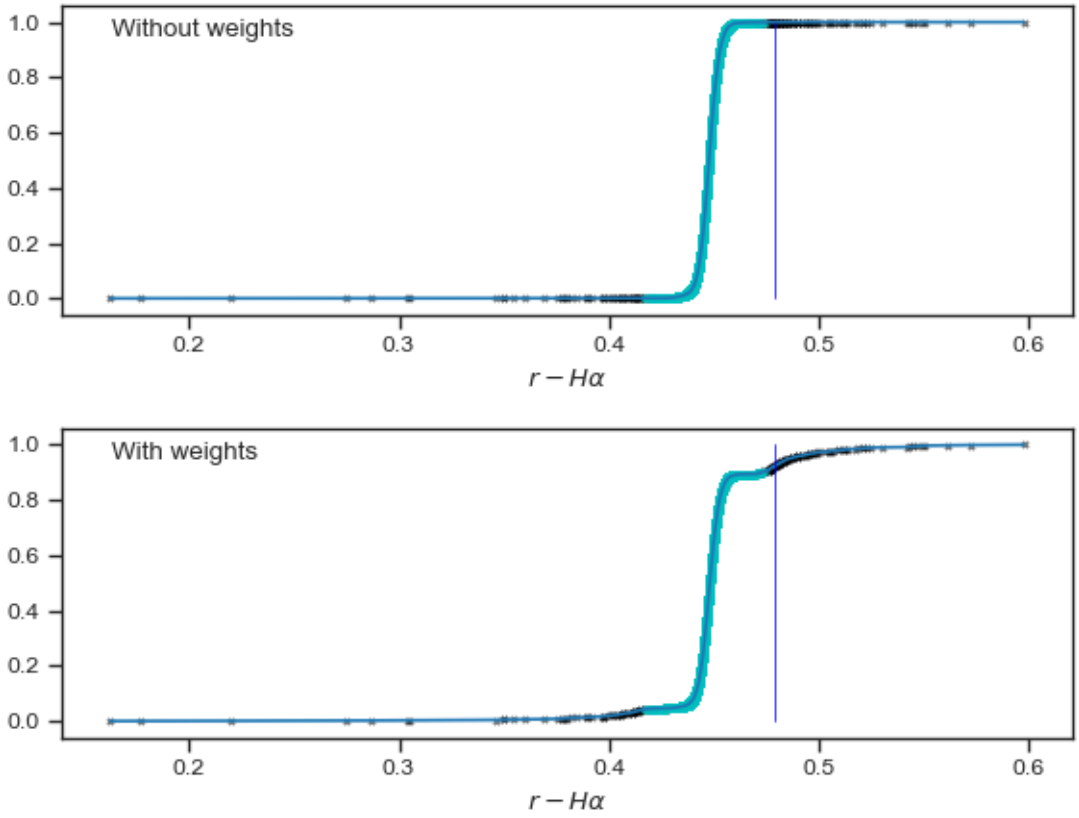


Figure 3.17: CDF plots for the SoI shown in figure 3.15, (*top*) without weights, standard EP; and (*bottom*) with weights and a threshold of 1000, weighted EP. Without weights all "noise" sources sit at either an EP score of 0.0 or 1.0, which means that variations, even large ones, in r-$H\alpha$ values are not represented by the EP score. Adding weights reduces this and variations are captured in the score with more extreme values having a score closer to 0.0 or 1.0.

Nearest cluster neighbour score/distance

The EP and its weighted variation, discussed previously, determine the r-$H\alpha$ deviation of a SoI with respect to the locus in its selection neighbourhood, but, this does not necessarily mean that an outlier based on the EP or WEP criteria is an actual emitter. Cases where the r-$H\alpha$ range of the selection neighbourhood could be limited to a small range close to the population or in the case of the standard EP, the large difference between number of cluster and "noise" sources would yield outliers that are most likely not $H\alpha$ emitters. To prevent this incorrect selection of emitters, the nearest cluster neighbour (NCN) score was considered, to be used in combination with the EP or WEP score. The NCN score is the euclidean distance from the SoI to the nearest cluster point.

Given the large number of data points, brute force calculation for this is not an option; there-

fore the kd-tree implemention from SciPy [55] was used to compute these values for all SoIs. The NCN score, by itself, is not suitable to select emitters as it does not take into account the distribution of the selection neighbourhood and is based upon arbitrary results of the clustering algorithm. The aim of the NCN is to ensure that all sources selected as emitters based on the EP or WEP score have a minimum distance from the closest locus. This does make the assumption that cluster sources, to some degree represent the closest locus, but given that the clustering is done based on density, this is an acceptable assumption to make.

k-Nearest cluster neighbour *r*-$H\alpha$ distance

The NCN score works well for all points that are directly above a locus (in the *r*-$H\alpha$ dimension), as the main contribution to the distance of the closest cluster source will come from the *r*-$H\alpha$ dimension. However, for other sources such as the one shown in figure 3.18, the distance is dominated by the *r*-*i* dimension; this results in inconsistent *r*-$H\alpha$ distances from the closest locus. Additionally, using only a single point to calculate this score can result in smaller than expected score values, as border points of the cluster might not be representative of where the cluster starts. To address these problems the NCN score was modified to the k-nearest cluster neighbour (kNCN) score, which is determined by finding the k nearest cluster neighbours (based on the euclidean distance) in the full three dimensional space and then calculating the mean *r*-$H\alpha$ distance.

For SoIs with their selection neighbourhood determined via slicing, section 3.7.2 a slight modification was done; the k-nearest cluster neighbour sources are determined based on a shifted *r*-*i* and M_G position. As the actual *r*-*i*, M_G position can be quite far next to the cluster, as for the SoI in figure 3.18, the selection of the k-nearest cluster sources does not necessarily represent the largest cluster *r*-$H\alpha$ values of the selection neighbourhood. Therefore the median *r*-*i* and M_G values of the cluster sources (in the selection neighbourhood) are used with the actual SoI's *r*-$H\alpha$ value to determine the kNCN score.

Increasing the number of neighbours considered should also increase the consistency and smoothness of this score, resulting in an improvement in the selection of the $H\alpha$ emitters. A value of $k = 5$ was used.

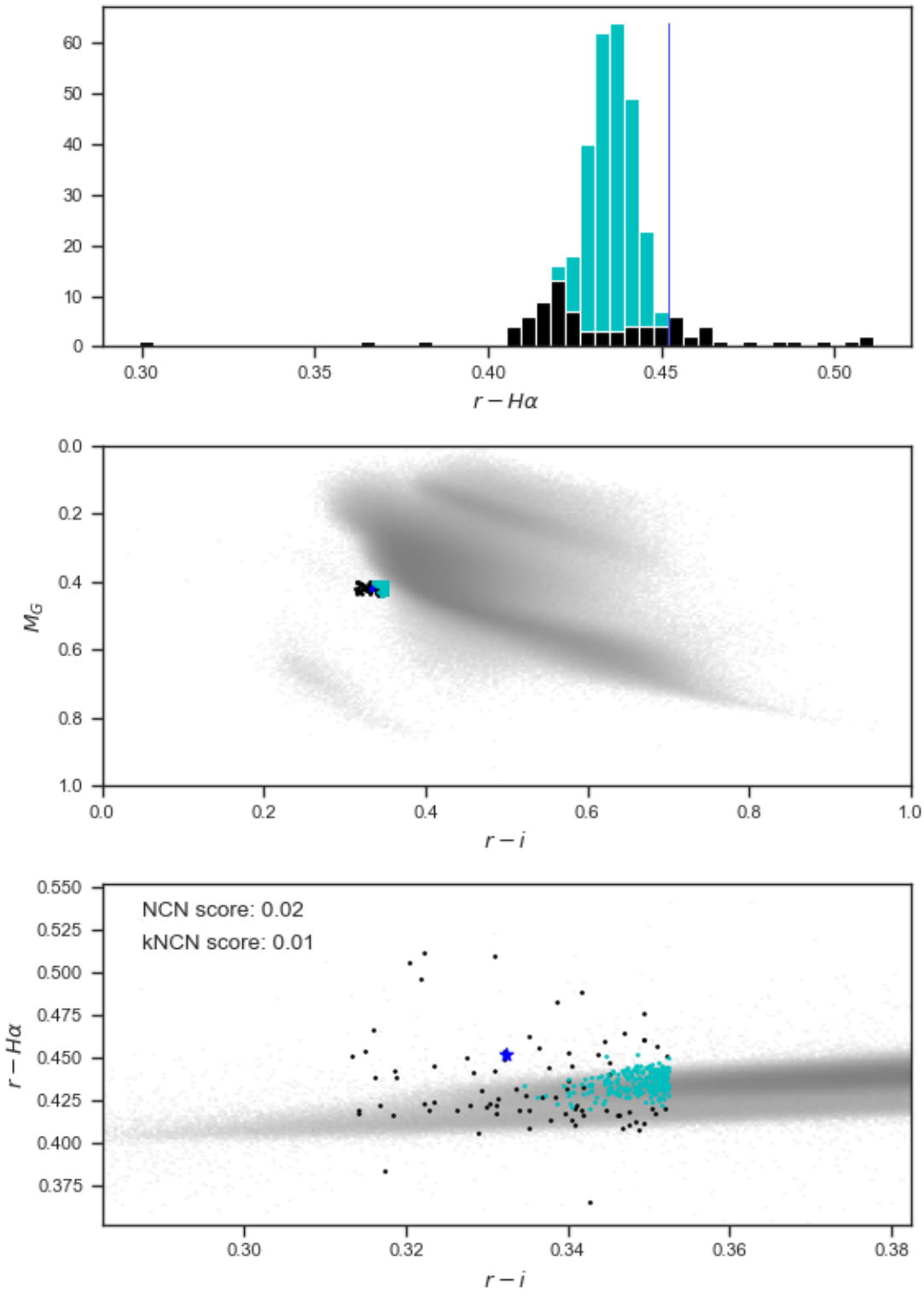


Figure 3.18: An SoI along on the edge of the main-sequence cluster is shown. The last plot indicates the SoI's position with respect to the selection neighbourhood and shows the NCN score is an unsuitable measure, in this case, for ensuring that potential emitter SoIs have a minimum distance in the r-$H\alpha$ dimension.

Chapter 4

Classification and Results

4.1 Overview

The diagram in figure 4.1 shows the full $H\alpha$ emitter selection process for all SoIs, i.e. all sources classified as noise by the DBSCAN (and NN) algorithms. A quick overview of the selection process is given here, with more details on the thresholds and type of SoIs given in the following sections of this chapter.

Tunnelling was performed for all sources that were classified as "noise" by the DBSCAN (and NN) algorithm. The next step was to determine whether or not the selection neighbourhood for a given SoI meets the "*on cluster*" condition covered in section 3.7.1 and represented by conditions **B-D** in figure 4.1, with sources reaching condition **E** in figure 4.1 having the "*on cluster*" flag set in the catalogue. Condition **E** then checks that the SoI has a greater r-$H\alpha$ value than all cluster sources in the selection neighbourhood, at which point the SoI is considered "*above cluster*". The scores discussed in section 3.7.4 are then calculated for all SoIs considered "*above cluster*" in step **4**.

All SoIs not meeting conditions **A-D** were therefore considered "*beside cluster*" and the selection neighbourhood was determined using slicing (step **2**, section 3.7.2), with scoring for these SoIs done in step **3**. SoIs not meeting condition **A**, of which there are only about 200, were scored (step **1**) using the selection neighbourhood of the closest reference point.

With scores calculated for all SoIs, emitters are then selected at condition **F** using the score thresholds defined in the following section 4.2. This is then followed by a brief discussion of the classification of the different type of SoIs.

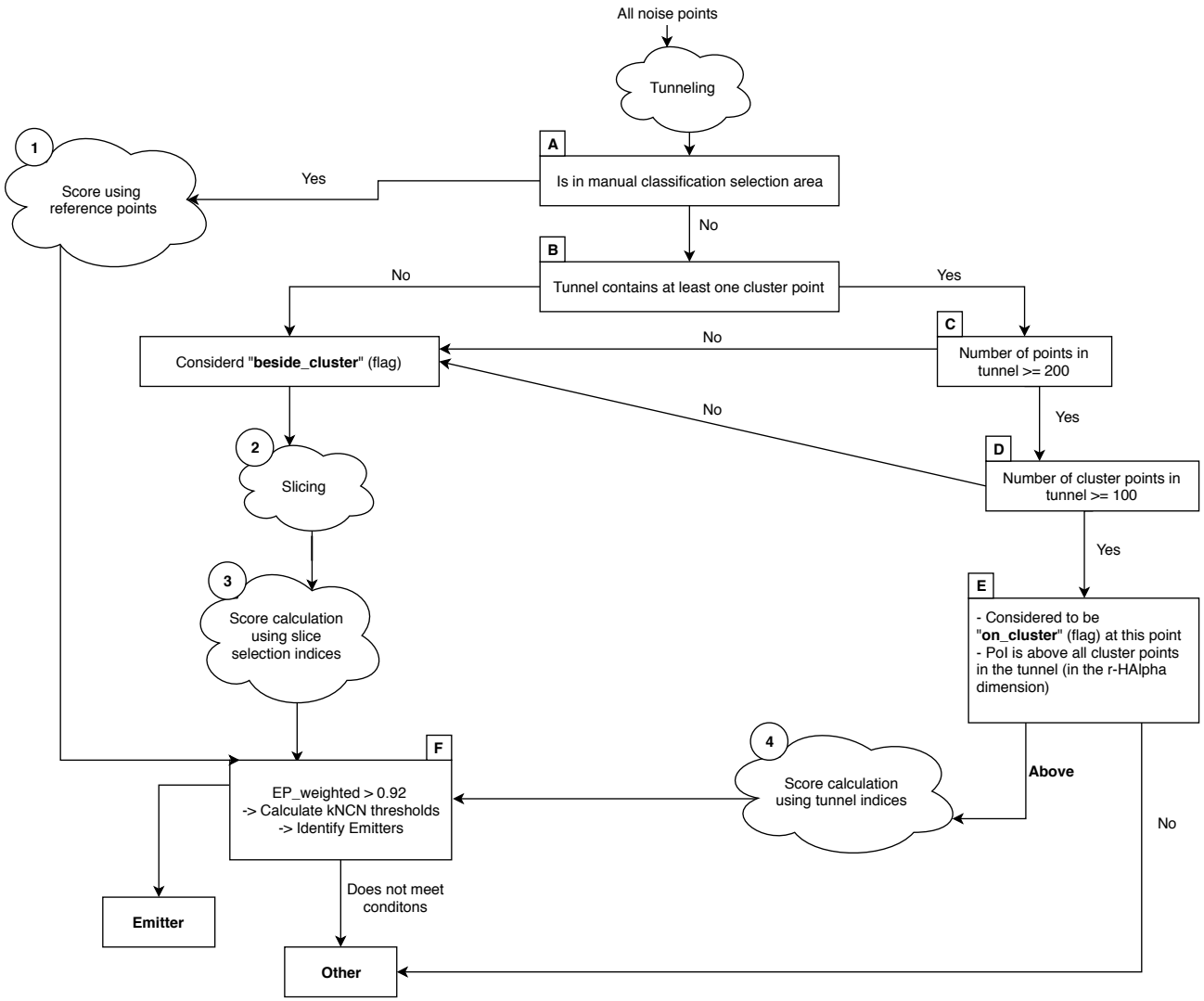


Figure 4.1: Classification diagram

4.2 Score Thresholds

The scores used for the selection were the WEP and kNCN scores discussed in section 3.7.4. Initially the EP and NCN scores were used to identify emitters. However, once their shortcomings (discussed in section 3.7.4) became apparent, these were modified to WEP and kNCN which were subsequently used for the selection of $H\alpha$ emitters.

Initially, constant thresholds were used for selection of emitters, such as $kNCN > 0.01$ and $WEP \geq 0.97$, as shown in figure 4.2. However, this resulted in missing emitters that were not in the top x-percentile of their selection neighbourhood. This effect is much larger for low-count selection neighbourhoods as the number of sources that can be in the top x-percentile is smaller. This can result in missing emitters, such as the one in figure 4.3. However, just decreasing the WEP threshold results in the pollution of the emitters with many points from the edge of the population meeting the kNCN threshold, especially for an SoI with a large selection neighbour-

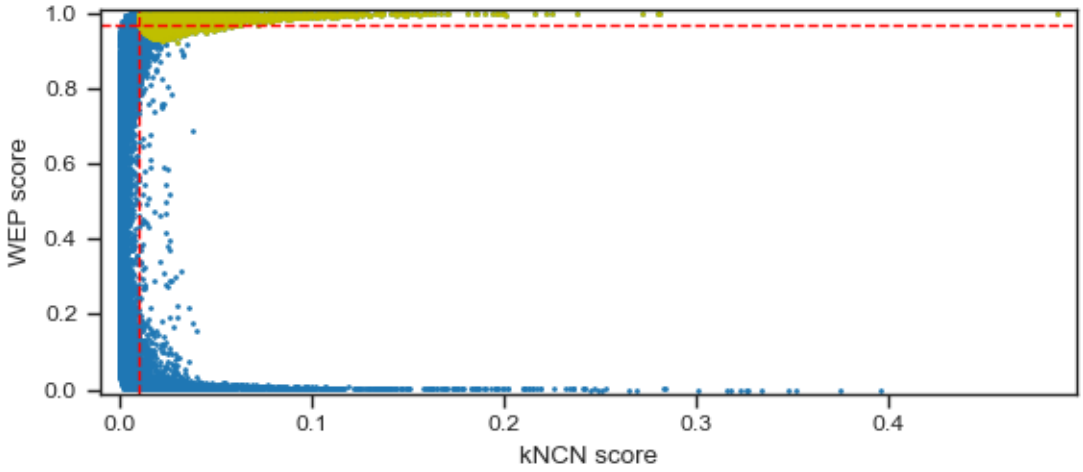


Figure 4.2: Shown are all SoI sources with the kNCN score along the x-axis and the WEP score along the y-axis. The red dashed lines represent an example of using constant thresholds to identify emitters. This method will miss a considerable number of emitters, assuming the WEP thresholds is set reasonably high to prevent the inclusion of non-emitters, as there will be SoIs that are not in the top x-percentile of their selection neighbourhood, as some CMD region have a larger amount of emission line objects. The emitters selected using a lower constant WEP threshold, and a WEP score dependent kNCN threshold are shown in yellow.

hood count, even when using the weighted EP score.

To account for this, the kNCN threshold was determined as a function of the WEP score, with the kNCN threshold increasing as the WEP score decreases as shown in figure 4.4. This approach allows reducing WEP threshold without the inclusion of SoIs that are most likely not emitters, therefore allowing the selection of SoIs with a lower WEP score but a large kNCN, which qualifies them as $H\alpha$ emitter objects. The calculation of the kNCN threshold is done as follows:

$$T_{raw}(x) = a^{(x-0.9)\mathrm{b}} \tag{4.1}$$

where the constants a and b are determined using trial and error using visual inspection of sources included/removed. The resulting thresholds are then scaled to the appropriate range using:

$$\begin{gathered} S = \{\text{all } T_{raw} \text{ values, for SoIs which have } WEP > 0.92\} \\ T_{scaled} = \frac{T_{raw} - \min_{x\in S} T_{raw}(x)}{\max_{x\in S} - \min_{x\in S}} \times (\mathrm{th}_{max} - \mathrm{th}_{min}) + \mathrm{th}_{min} \end{gathered} \tag{4.2}$$

where th_{min} and th_{max} are the user defined min/max thresholds. These determine the kNCN thresholds at the WEP extremes, i.e. at $WEP = 1.0$ the corresponding kNCN threshold is th_{min} and at $WEP = 0.92$ the kNCN threshold is th_{max}. The WEP threshold used was 0.92 with the used kNCN threshold shown in figure 4.4.

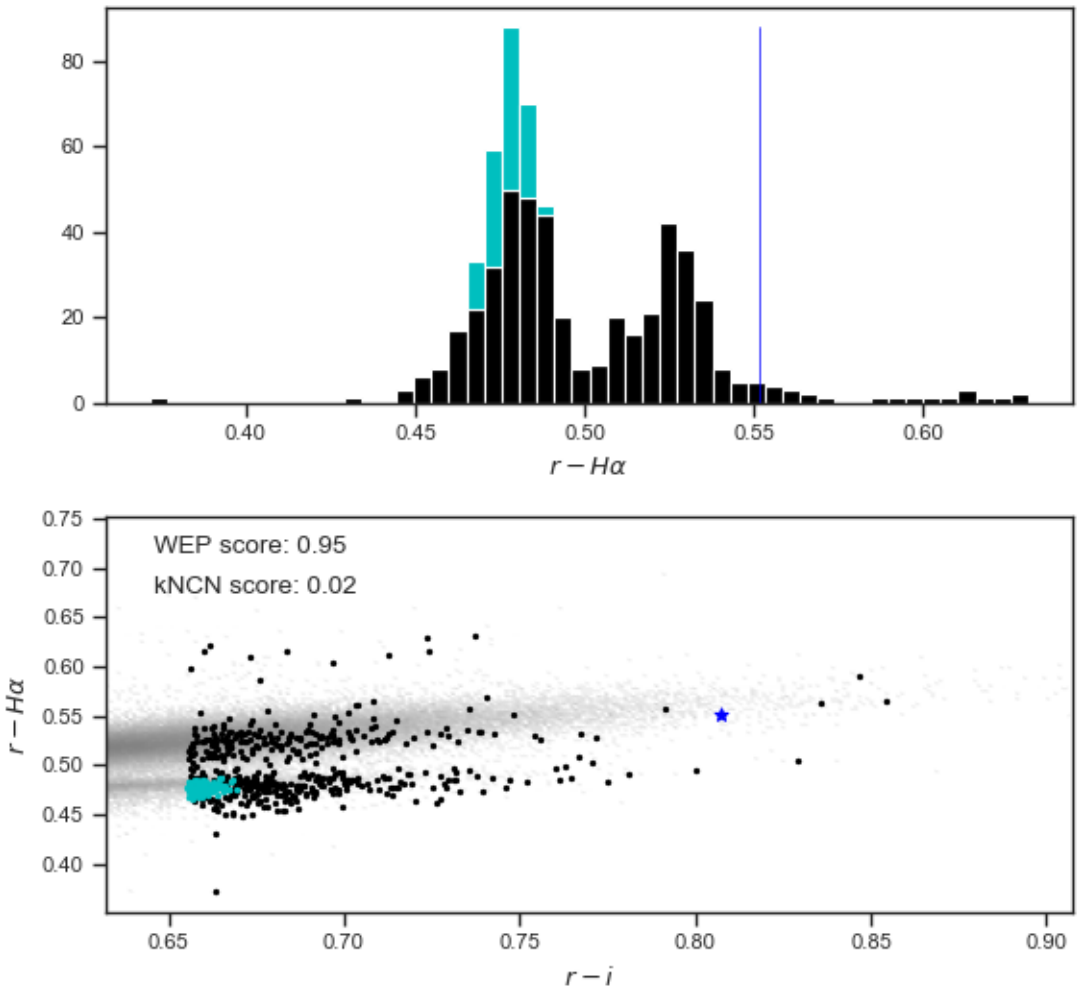


Figure 4.3: This is an example of an SoI which is most likely an $H\alpha$ emitter, but would have been missed if a constant kNCN threshold had been used. Using a constant kNCN threshold requires a higher WEP threshold to prevent selection of sources that are likely not emitters.

4.3 Selection for Different SoIs Types

4.3.1 On Cluster

Out of the 21,381 SoIs, 15,523 are considered "*on cluster*" as defined in section 3.7.1 or as determined according to the diagram in figure 4.1. These points can be found using the "on_cluster" flag in the catalogue. In order to qualify as "*above cluster*", a SoI has to have a *r-Hα* value greater than all cluster sources in its selection neighbourhood, represented by condition **E** in the diagram. It is worth noting that to be "*on cluster*", the SoI itself does not actually have to be directly on the cluster, rather the tunnel has to contain the required number of cluster points and overall sources, and given that the tunnel has a certain width/height in the CMD

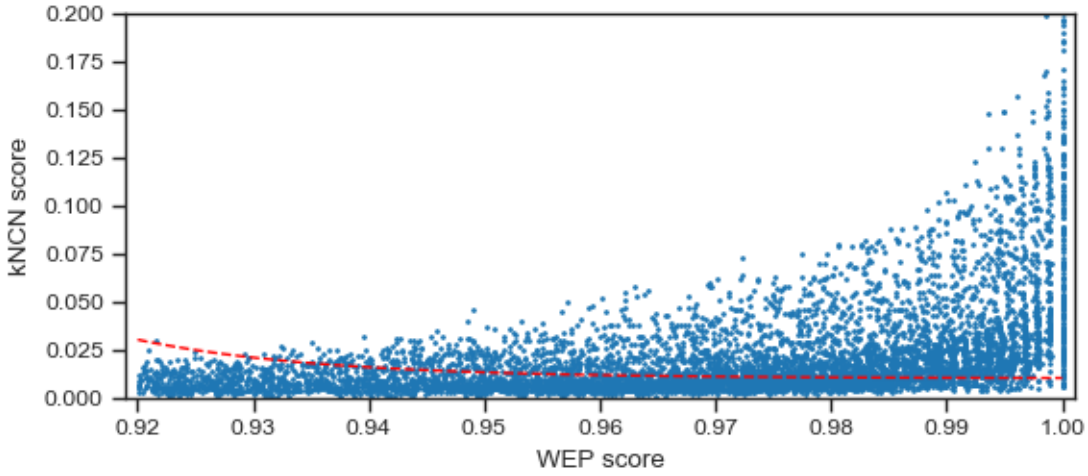


Figure 4.4: Shown are all SoIs with $WEP \geq 0.92$ and the kNCN threshold as a function of WEP.

plane SoIs close to a cluster can also be classified as "*on cluster*" given that all the conditions are meet. However, in general, SoIs meeting the conditions **B-E** are generally "*on cluster*" in the CMD plane, hence the naming. Sources that are not actually directly above a cluster are discussed further in section 4.3.3.

Only a small number of SoIs above the white dwarf population are considered to be "*on cluster*", which is due to its low density and hence the count and cluster count thresholds (section 3.7.1) not being met. Modifying the thresholds to allow inclusion of more SoIs as "*on cluster*" for the white dwarf population would result in reduction of sources in a selection neighbourhood, which would lead to an increased incorrect selection of SoIs as emitters. Therefore most points on the white dwarf population are processed using the *beside cluster* approach, covered in section 4.3.4 and 3.7.2; also shown in the left branch of the diagram in figure 4.1.

4.3.2 Above/Below Cluster

Out of the 12,523 "*on cluster*" SoIs, 4,847 are considered above cluster, i.e. meeting conditions **B-E** in the diagram in figure 4.1. These conditions ensure that the tunnel area is above a cluster in the CMD plane and contains enough points in its selection neighbourhood to allow selection of $H\alpha$ emitters using the WEP score. Selection of $H\alpha$ emitters from this group of SoIs is done by calculating the WEP and kNCN score with respect to their respective selection neighbourhoods; these scores are then checked against the thresholds defined in section 4.2. Out of the 4,847 SoIs considered "*above cluster*", 3097 were selected as $H\alpha$ emitters.

4.3.3 Border Points

This is a small subgroup of the SoIs considered "*above*" cluster and are not scored or treated any different. These are SoIs which are not actually directly above or on a cluster in the CMD, but instead sit on the border of a cluster, and their tunnels include enough of the cluster sources to meet the required conditions to be considered above a cluster, which means that the tunnel is used as the selection neighbourhood. Since the cluster is defined by arbitrary parameters and only represents the local population to a certain degree, including these sources makes sense.
The main difference is that these SoIs have a much smaller number of sources in their selection neighbourhoods compared to non-border points considered "*above cluster*" and the proportion of other "noise" sources is much larger. This can be seen by in figure 3.11, where the *Center* and *Center-Bottom* plot show the selection neighbourhoods *r*-$H\alpha$ distribution of two border points and the *Bottom* plot shows a "*above cluster*" non-border point. However, this makes no difference as the whole idea is to select emitters with respect to the selection neighbourhood which is part of the local locus, so the cluster count condition purely exists to ensure that a core "part" of the locus is included. The "noise" sources that are included in the tunnel of a border point are as much part of the locus as the cluster points; hence the main difference is the lower number of sources and the distribution looks different in terms of cluster to "noise" proportion. This only really applies to the main sequence and giant cluster; as the white dwarf locus has a much lower density, hence all points considered "*above*" the white dwarf cluster are actually directly above the cluster.

4.3.4 Beside Cluster

All SoIs (with the exception of the small number of manually classified SoIs covered in section 4.3.5) not considered "*on cluster*" are denoted "*beside cluster*" and can be found using the "*beside_cluster*" flag in the catalogue, of which there are 5,629.
For these SoIs, tunnelling did not produce a suitable selection neighbourhood as not all conditions were met (as discussed in section 3.7.1, or shown in the digram in figure 4.1 as conditions **B-E**). To further explain why tunnelling is unsuitable, figure 4.5 shows all SoIs that are considered "*beside cluster*". This figure shows nicely that these sources are all in very low density regions of the CMD. This means that using tunnelling for these SoIs would result in almost empty selection neighbourhoods, making selection of $H\alpha$ emitters inaccurate. In addition, the selection neighbourhoods would most likely not include the population the SoI belongs to. Therefore slicing (explained in section 3.7.2 and shown as step **2** in the diagram) is used.
As with all other SoIs, the scores are calculated and compared against the thresholds covered

in section 4.2 and $H\alpha$ emitters are selected accordingly.

Using slicing to determine the selection neighbourhood allows the scoring and selection to be done in the same manner as for SoIs in dense regions, which keeps it consistent. However, it is worth noting that this also means that SoIs beside a cluster are only ever compared against the edges of the population along with the "noise" sources close to the border. Additionally, for some SoIs, neither of the two slices intersects with their closest population, as shown in figure 3.10. These SoIs are further discussed in section 4.3.5.

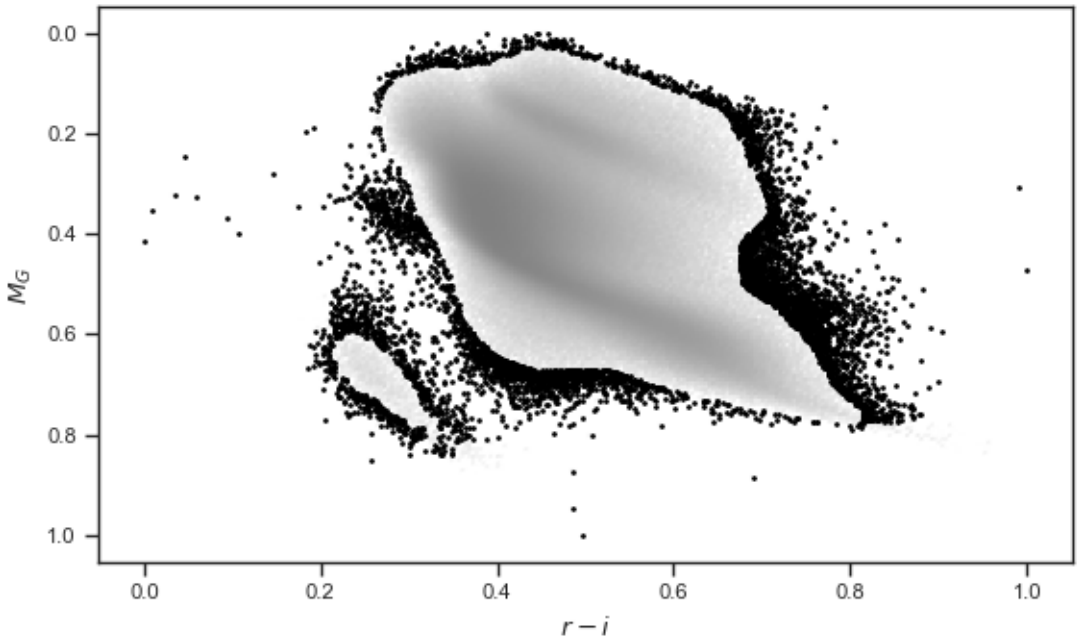


Figure 4.5: All sources considered "*beside cluster*" are shown in black on a colour-magnitude diagram.

4.3.5 Manual Classification

Slicing is not suitable for all "noise" sources that are considered beside a cluster, as in some cases neither of the two slices intersects with the closest locus. These sources are shown in figure 3.10 in yellow. One approach to fixing this problem would be to add more slices, or use the closest sources instead of slices to determine the selection neighbourhood. However, given that most SoIs are either above or below a cluster and hence do not require slicing, and the number of SoIs for which slicing or tunnelling is not an option is very small ($\sim$200), this approach was used to avoid having to add more complexity. These sources are selected by the blue lines as shown in figure 3.10 which displays all SoIs that required manual classification. The selection neighbourhood is then determined from a close manually-selected reference point by tunnelling

with manually configured width/height for each manual classification area. With the selection neighbourhood determined, the scores are then calculated and, as with the other types of SoIs, these scores are then checked against the thresholds, allowing selection of $H\alpha$ emitters.

Chapter 5

Validation

In order to determine how well the selection algorithm covered in the previous sections performs; validation is done manually with plots, comparison with the Witham et al. (2008) [66] $H\alpha$ emitters catalogue and crossmatching with SIMBAD and LAMOST.

5.1 Manual Validation

Manual validation was done by gridding the dataset with a much lower number of cells. From each cell an emitter and non-emitter SoI was selected. For each of the selected sources, plots were created depending on how their selection neighbourhood was determined, i.e. either slicing or tunnelling. The plots, along with the SoIs score, were then examined manually to identify any issues with the selection algorithm and gauge its accuracy based on the three dimensional data available. To completely confirm that a specific SoI is an $H\alpha$ emitter, spectroscopic follow up observations have to be done, but this was not done as part of this work.

Examples of the plots used are shown in figures 5.1 and 5.2. Both of the two different plot types show the SoI's position on the CMD, along with a histogram of the r-$H\alpha$ values of the selection neighbourhood sources, which were used to score the SoI.

Manual validation showed that overall the selection algorithm works well. Investigating about 200 SoIs using these type of plots (figures 5.1 and 5.2), showed that only a very small fraction of selected $H\alpha$ emitters were selected incorrectly. However, without spectroscopic follow-up observations it is difficult to be certain if they are actual $H\alpha$ emitters or not. Manually inspecting these plots also showed that the thresholds used are missing some $H\alpha$ emitters. However, given that a clean sample of $H\alpha$ emitters was prioritised over a more complete sample, this is to be expected. Furthermore some limitations in the selection algorithm also became apparent. These are discussed in section 6.3.

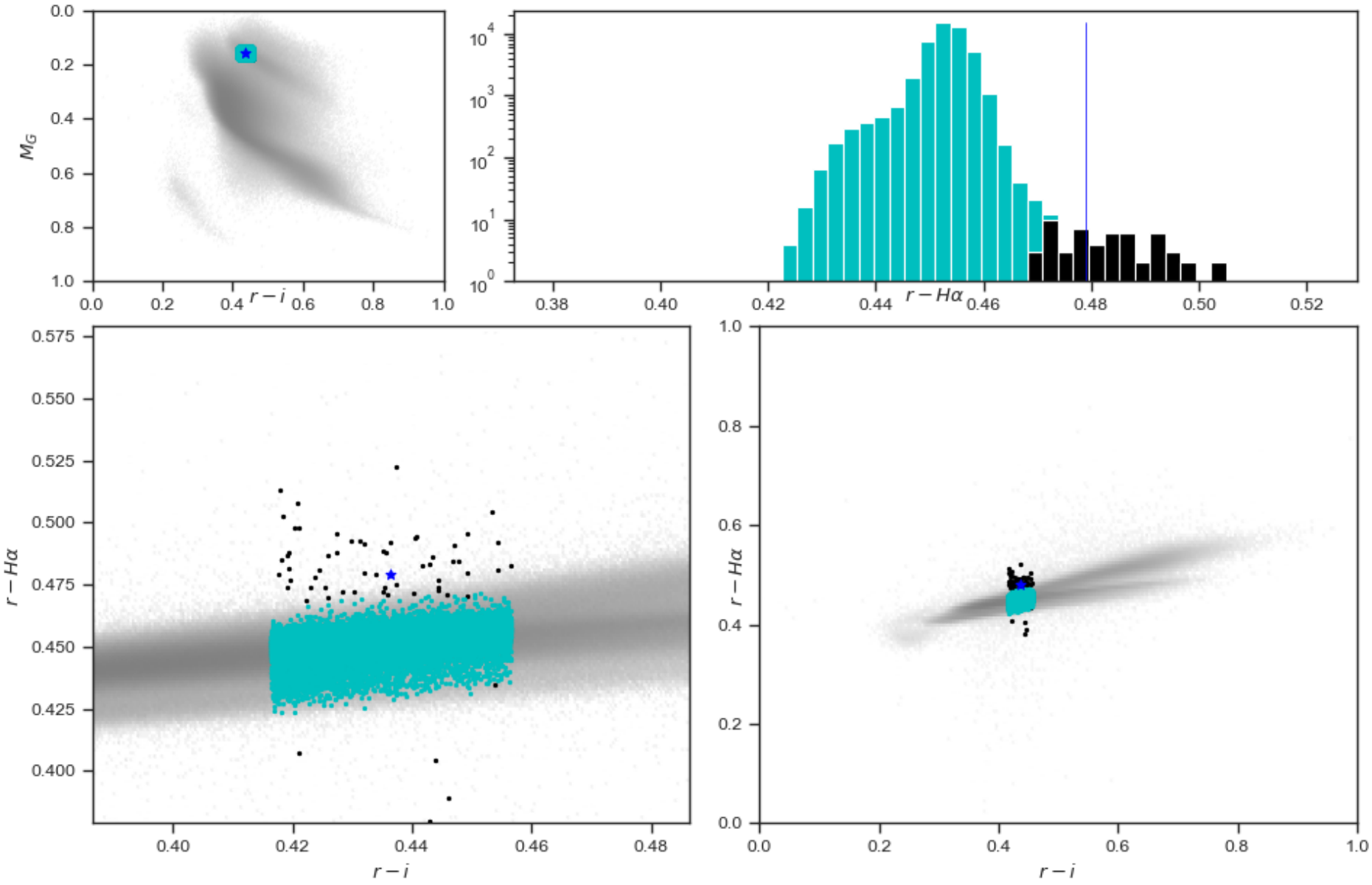


Figure 5.1: Manual validation example plot for an "*above cluster*" SoI, i.e. its selection neighbourhood selected via tunnelling (section 3.7.1).

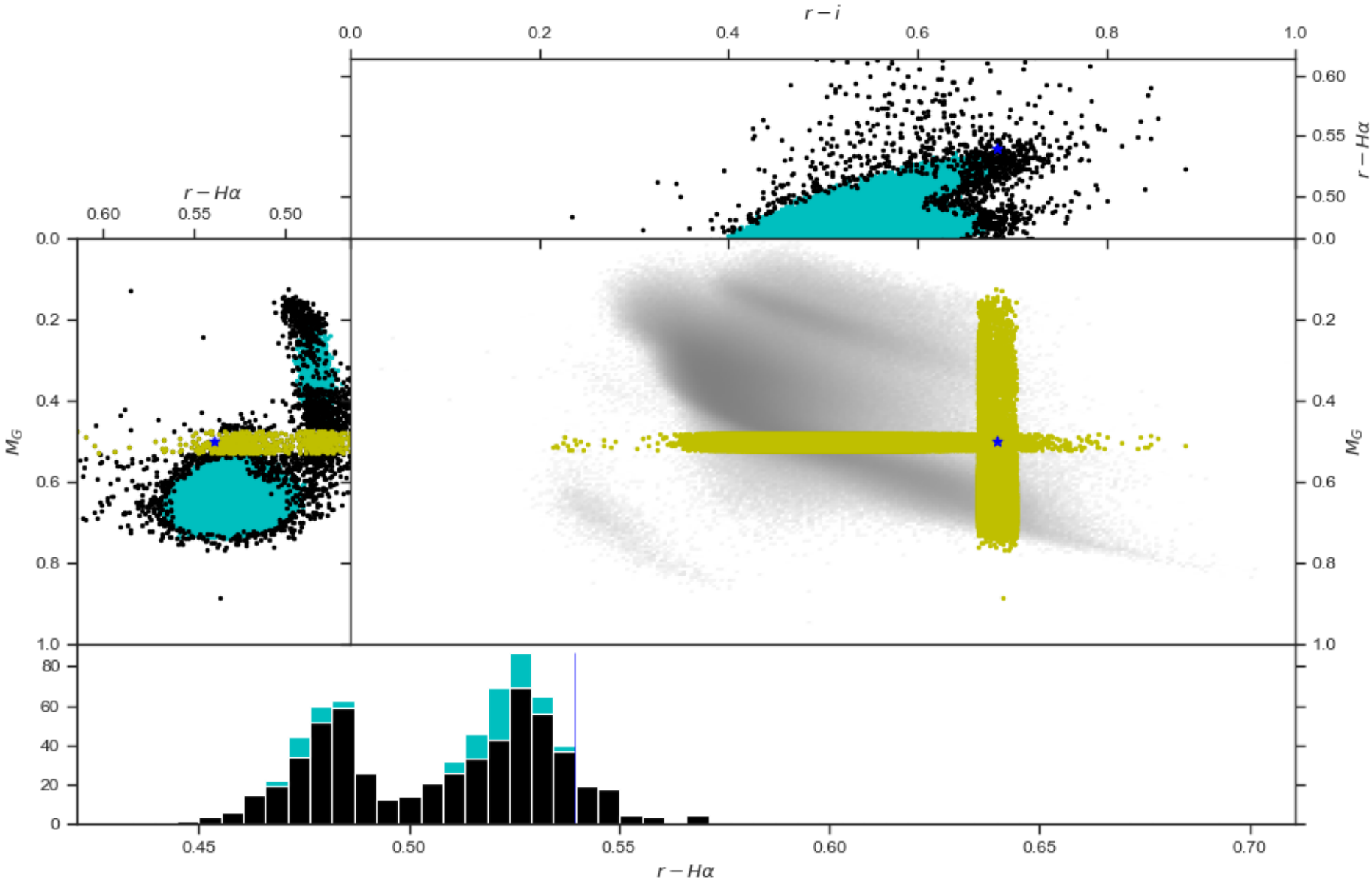


Figure 5.2: Manual validation example plot for a "*beside cluster*" SoI, meaning its selection neighbourhood was selection using slicing (section 3.7.2).

5.2 LAMOST

The Large sky Area Multi-Object fibre Spectroscopic Telescope (LAMOST) is conducting a spectral survey in the northern sky [69]. Crossmatching the identified emitters with the LAMOST DR4 using the Vizier website [62] gives access to the spectra of the sources that were matched. This allows further validation for how well the selection algorithm performs. As downloading the spectra was not straightforward, the validation was done by manually inspecting the spectra on Vizier.

During this process it was identified that some of the selected $H\alpha$ emitters were not actually emitters and the large $H\alpha$ is due to saturation. Therefore the selection algorithm was run again, with the initial saturation cuts $r > 13$, $i > 12$ and $Ha > 12.5$, as suggested by IPHAS [6] (Table 1), adjusted to $r > 13.5$, $i > 12.5$ and $Ha > 13$. Rerunning the selection algorithm with the new cuts, resulted in the removal of 433 sources classified as emitters, of which 407 were removed directly due to the new cuts. Overall the new cuts removed ∼100,000 sources, which most likely explains the 26 previously-selected emitters no longer being selected as emitters. Checking the 407 sources that exceed the initial saturation limits against known SIMBAD sources, shows that 22 classified as Be stars (Be*) and 23 as Emission line stars (Em*). Crossmatching the 407 removed sources with the LAMOST DR4 using Vizier and a radius of 2 arc seconds resulted in 126 matches. The spectrum of every fourth star was manually viewed and classed as either an emitter or non-emitter based on the spectrum; a reasonably straightforward process as the example spectrum in figure 5.3 shows. Out of the 32 spectra viewed, 21 did not show any emission at the $H\alpha$ wavelength while 10 did. Given that all of these sources were classified as emitters by the selection algorithm, this highlights that there is a problem with the bright sources, when compared to the LAMOST validation groups which have a much higher proportion of LAMOST spectroscopically-confirmed emitters.

The selected emitters were split into six different groups based on their WEP and kNCN scores, and for each group ∼30-40 spectra were manually viewed, unless the number of crossmatches was lower in which case all available spectra were viewed. The crossmatching was performed with a radius of 2 arc-seconds. The six different groups used were:

A) $WEP \geq 0.98$ and $kNCN \geq 0.05$

B) $WEP \geq 0.98$ and $kNCN \leq 0.05$

C) $0.98 \geq WEP \geq 0.94$ and $kNCN \geq 0.05$

D) $0.98 \geq WEP \geq 0.94$ and $kNCN \leq 0.05$

E) $0.94 \geq WEP \geq 0.92$ and $kNCN \geq 0.05$

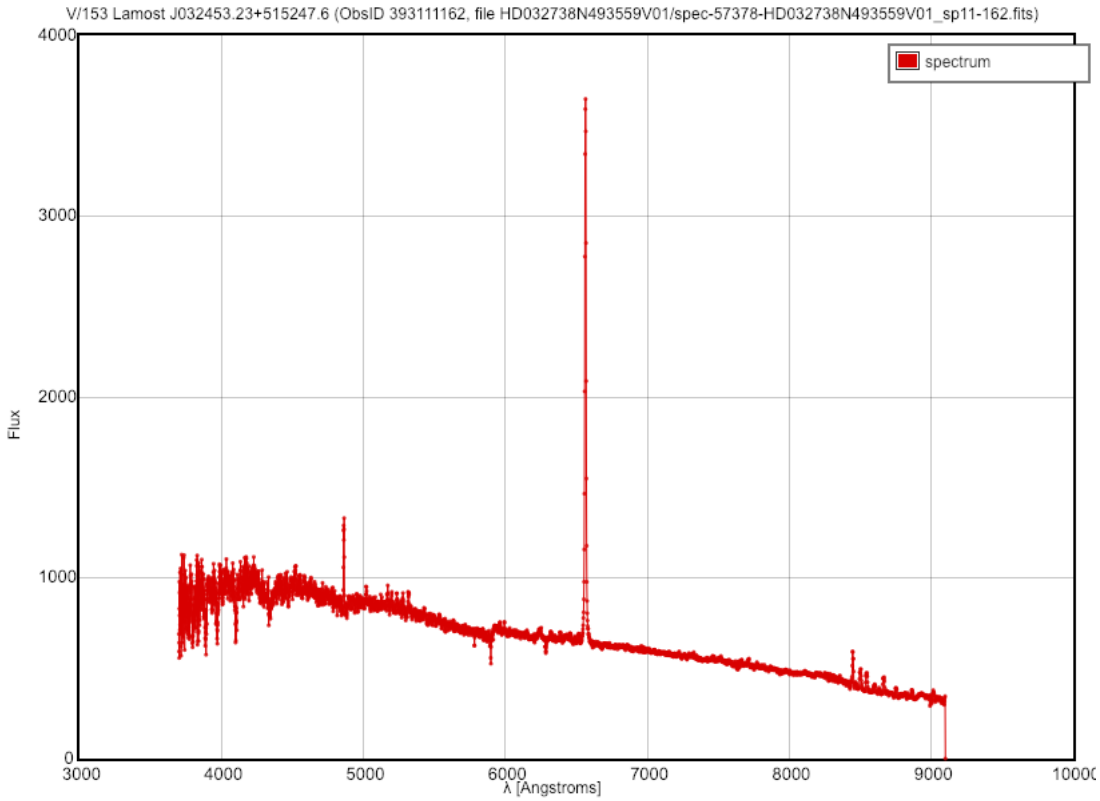


Figure 5.3: Example LAMOST spectrum of a selected $H\alpha$ emitter.

F) $0.94 \geq WEP \geq 0.92$ and $kNCN \leq 0.05$

The emitters were split into the groups to see how the ratio of correctly-selected emitters versus selected non-emitter sources depends on the WEP and kNCN scores. Some matched sources had more than one spectrum, but these agreed in all cases and were counted as one match. If a source matched with more than one target entry from LAMOST, both were viewed and if one of them was an emitter then this was counted as an emitter match; this case only occurred once. Table 5.2 shows, that while the number of matches is limited with LAMOST, the accuracy of the selection algorithm is good, with an accuracy of ∼90%. There is an increase in the number of incorrect selections in the lower score groups, but this is to be expected and is a further validation that the WEP and kNCN score is a good combination for the selection of emitters. The second to last row in table 5.2 shows the effect of saturation on the selection algorithm, with 21 incorrect emitter selections to the ten correct ones. That corresponds to a mis-classification rate of 48%, clearly indicating that there is an issue with the sources between the new and original saturation cuts. To further highlight the increase in the of number incorrect selections as the thresholds are decreased, the last row in table 5.2 shows the number of incorrect selections for the additional selected emitters if the thresholds were modified as shown in figure 5.4. The number of misclassifications goes from $\sim 9\%$ to $\sim 33\%$ based on the LAMOST validation. This

Catalogue		LAMOST		
Group	Entries	Matches viewed	Emitters	Non-Emitter
A	576	19	19	0
B	1519	36	34	2
C	59	2	2	0
D	1236	37	31	6
E	0	-	-	-
F	57	2	1	1
Totals	**3447**	**96**	**87**	**9**
Other				
Emitters removed by adjusted cuts	407	32	10	21
Modified thresholds as per figure 5.4	1737	45	30	15

Table 5.1: Table showing the results of LAMOST validation by splitting the selected emitters into groups and crossmatching with LAMOST DR4 using a radius of 2 arc-seconds. The last row shows the validation results for the bright emitters that were removed by the adjusted saturation cuts. The last row in the table shows the effect of reducing the threshold (as shown in figure 5.4) on the number of non-emitters polluting the selected emitters.

also highlights that the selected emitters in this catalogue are clearly not a complete or near complete set of $H\alpha$ emitters; there are many other $H\alpha$ emitters present in this dataset. However, given that the main aim of this catalogue is to assist with target selection for spectroscopic follow-up observations it was deemed more important to have a clean set of emitters versus a more complete one.

5.3 Simbad

SIMBAD (Set of Identifications, Measurements and Bibliography for Astronomical Data) is an astronomical database containing information on about 4,500,000 stars and 3,500,000 non-stellar objects (galaxies, planetary nebulae, clusters, novae etc.) with basic information such coordinates, proper motion, redshifts etc. for each object. The purpose of SIMBAD is to provide information on astronomical objects that have been studied in past scientific articles. In addition to the basic information, it also stores the type of each astronomical object as identified in a past study. Each object has a primary object type and a number of additional types, e.g. a studied Be star will have a primary type of Be* and the additional types Em* (emission line star) and * (star). A list of the different object types can be found here [59], along with general information

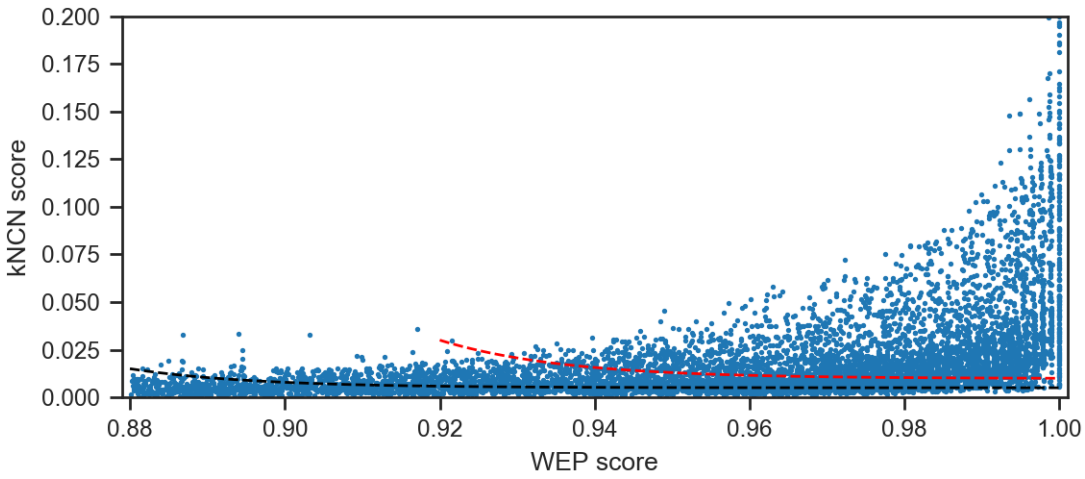


Figure 5.4: All SoIs with WEP score above 0.88, with the used kNCN threshold shown in red and the modified one, to show the effect of reducing it, shown in black.

on SIMBAD here [58]. Additionally this dataset of studied objects is also crossmatched with a large number of other catalogues and sources allowing easy lookup of information.

Gaia DR2 is a catalogue that has been crossmatched with SIMBAD, allowing simple lookup of SoIs using their respective Gaia DR2 identifier. Out of the 21,381 SoIs, 1,437 matches were found in the SIMBAD database, providing a good dataset for validation and analysis of the $H\alpha$ selection algorithm.

Table 5.3 shows the primary object type counts, additional object type counts and how many of those are classified as an emission line star. The number in front is always the number of selected emitters (by the selection algorithm described here) for the type in question. (To prevent confusion, sources classified as emitters by SIMBAD will be referred to as emission line objects, whereas the sources selected by the selection algorithm presented here will be referred to as emitters or $H\alpha$ emitters.) Out of the 1,437 matches with SIMBAD, 402 had the emission line star object type, which is just another word for $H\alpha$ emitter. This provides a direct measure of how well the selection algorithm worked. Out of the 407 SIMBAD emission line objects, 270 were correctly identified as $H\alpha$ emitters, i.e. 66%. This is a rather low percentage and further indicates that the completeness of the set of selected emitters has room for improvement. To get a better understanding of why the selection algorithm missed such a large number of emission line objects, manual validation plots (as discussed in section 5.1) were produced for these sources and then investigated manually.

Manually investigating these missed emitters, showed several situations in which the selection algorithm (with the threshold specified in section 4.2) is unable to successfully identify $H\alpha$ emitters. The main reason, explaining roughly 65 of the 135 missed SIMBAD emitters, is due

SIMBAD object type	Primary count	Total number with object class	Number of Em* objects
Emission-line Star (Em*)	148/201	270/405	270/405
Young Stellar Object (Y*O)	172/438	276/591	96/146
Young Stellar Object Candidate (Y*?)	43/83	171/277	56/74
T Tau-type Star (TT*)	109/206	109/206	71/129
T Tau star Candidate (TT?)	1/1	60/67	58/65
Variable Star (V*)	5/7	105/164	64/88
Be Star (Be*)	43/68	43/68	11/11
Variable Star of Orion Type (Or*)	16/18	48/64	41/50
Star in Cluster (*iC)	4/42	107/255	62/124

Table 5.2: The first column is the SIMBAD object type in question. The second column shows the number of objects that have this type as their primary type with the number in front showing how many of those have been selected as $H\alpha$ emitters by the selection algorithm. The third column shows the total number of the matched sources that have this object type and the number of selected emitters. The last columns shows how many of those objects are also classified as emission line star (Em*) by SIMBAD, where the number in front is the number of these emission line objects that were also selected as emitters by the selection algorithm. Note: not all object types of the matched sources are shown, only the ones with a reasonably large number of occurrences.

to the conservative selection thresholds used. An example is shown in figure 5.8. Modifying the thresholds, as shown in figure 5.4, allows an extra 55 of the 135 missed SIMBAD emitters to be selected by the selection algorithm. However, as table 5.2 shows, this also significantly increases the number of incorrectly identified emitters.

The second situation are emission line objects that are below their respective neighbourhood population on a colour-colour plot, most likely due to incorrect IPHAS values. An example is shown in the second panel of figure 5.8. These points are not identifiable as emitters with the dataset used, manually or with an algorithm.

The last scenario in which the selection algorithm fails to identify $H\alpha$ emitters correctly is when the selection neighbourhood contains cluster sources from more than one locus, an example of this is shown in the lower two panels of figure 5.8. The emitter in the "Dual population 1" panel of figure 5.8 was missed from an algorithm standpoint due to the few cluster points (cyan) to the left of the source of interest, resulting in a small kNCN value, meaning that the source will not be selected as an emitter. The SoI most likely belongs to the population in the lower, main locus of the selection neighbourhood, which would make it an obvious $H\alpha$ emitter. However, the small number of, most likely unreddened, cluster sources to the left of the SoI prevent the

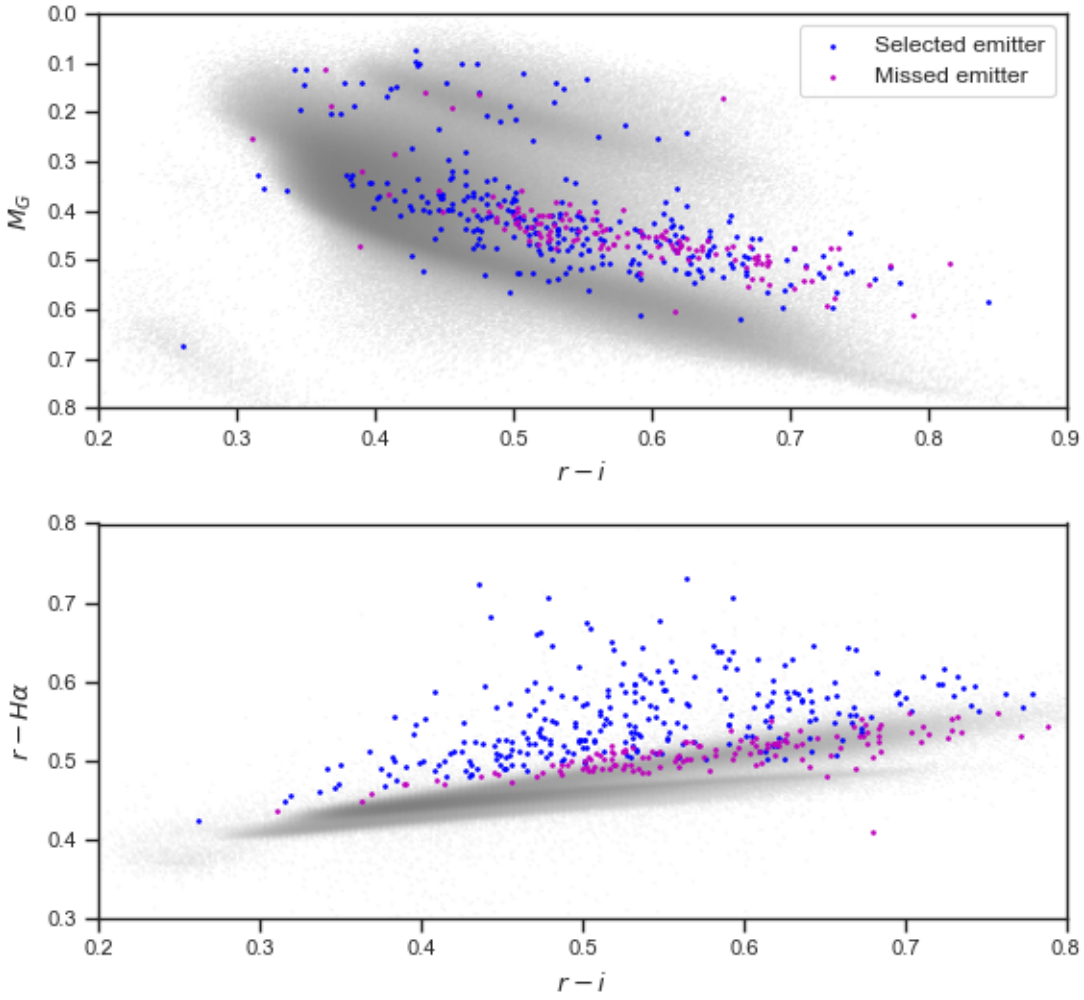


Figure 5.5: Colour-magnitude and colour-colour plot of all emission line star objects, with the colouring showing which ones the selection algorithm identified (blue). The missed emitters (purple) appear to be generally clustered in the tail end of the main sequence and none are obvious emitters that can be identified by just examinig the colour-colour plot.

selection algorithm from finding this $H\alpha$ emitter. As discussed in section 1.3 one of the aims of this project was to be able to identify these types of sources, as the absolute magnitude allows reasonably accurate (not adjusted for reddening or extinction) placement of sources on the CMD. However, while this worked to some degree (see results section 6.2) there are clearly scenarios in which the selection algorithm is not good enough to identify these emitters. There is also a similar of this scenario in which the selection algorithm is unable to identify the $H\alpha$ emitter, which is shown in the bottom panel of figure 5.8. The emitter most likely belongs to the red giants/reddened main-sequence locus. However, in this region, the two branches are only starting to diverge and some unreddened main-sequences sources are also in the selection neighbourhood, preventing the algorithm from finding this emitter.

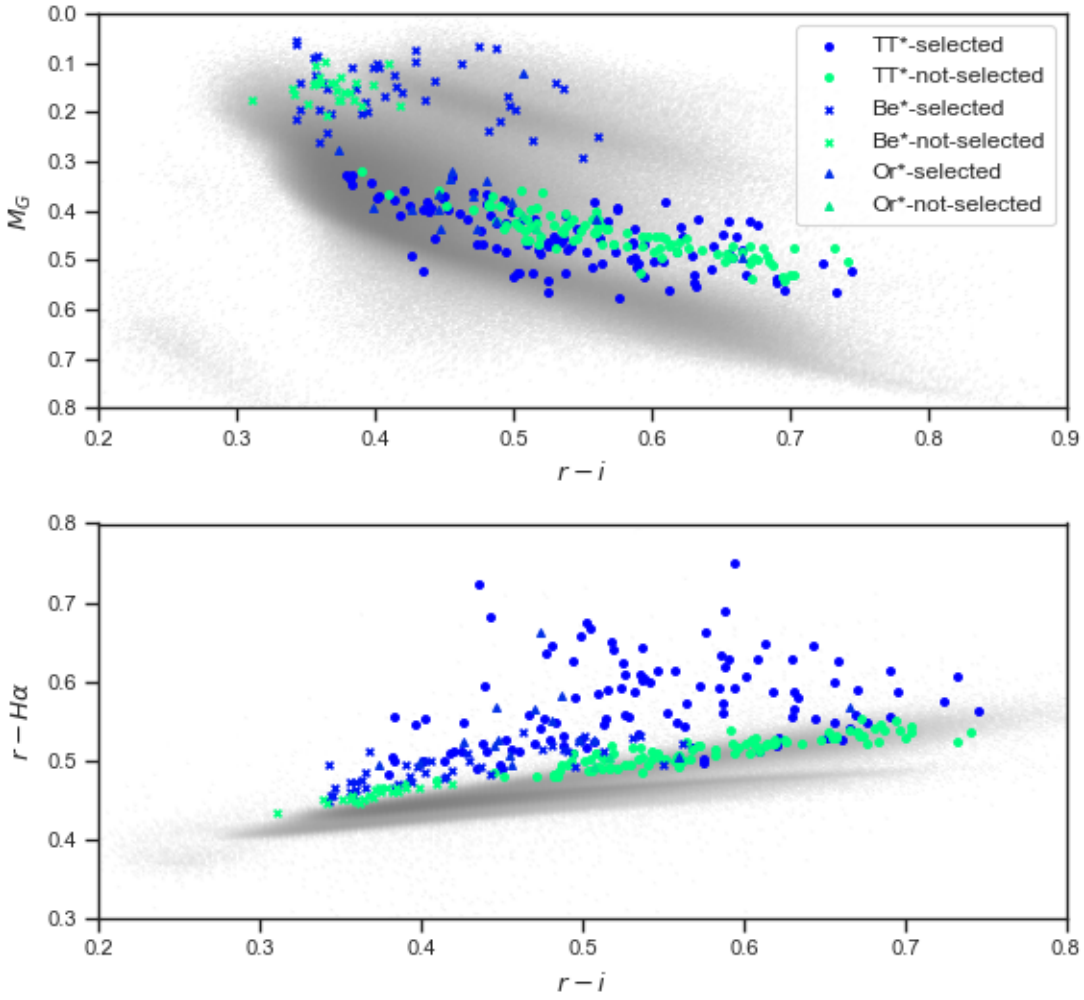


Figure 5.6: Colour-magnitude and colour-colour plot of the SIMBAD objects, except YSOs, with the colour indicating if it has been selected by the selection algorithm (blue) or missed (green).

5.4 Witham Comparison

Crossmatching the Gaia/IPHAS VAC with the Witham et al. (2008) catalog [66] using a 2 arc-second radius resulted in 1,232 matches; the number of matches is quite low given that both catalogues share IPHAS. However, this is most likely due to the different data cuts used by Scaringi et al. (2018) (covered in section 2 of [51]). Of the 1,232 crossmatches, 874 were selected as $H\alpha$ emitters by the selection algorithm and 358 were not classified as emitters. Some of the sources that were classified as emitters by Witham et al. (2018) [66] but not by this selection algorithm are possibly due to re-calibration in the second data release of IPHAS, used in the Gaia/IPHAS VAC catalogue. These differences are shown in figure 5.9 for all "missed" emitters. The blue triangles indicating the position based on the data from Witham et al. (2008) [66] and the black markers show the position based on the Gaia/IPAHS VAC catalogue. Many of the

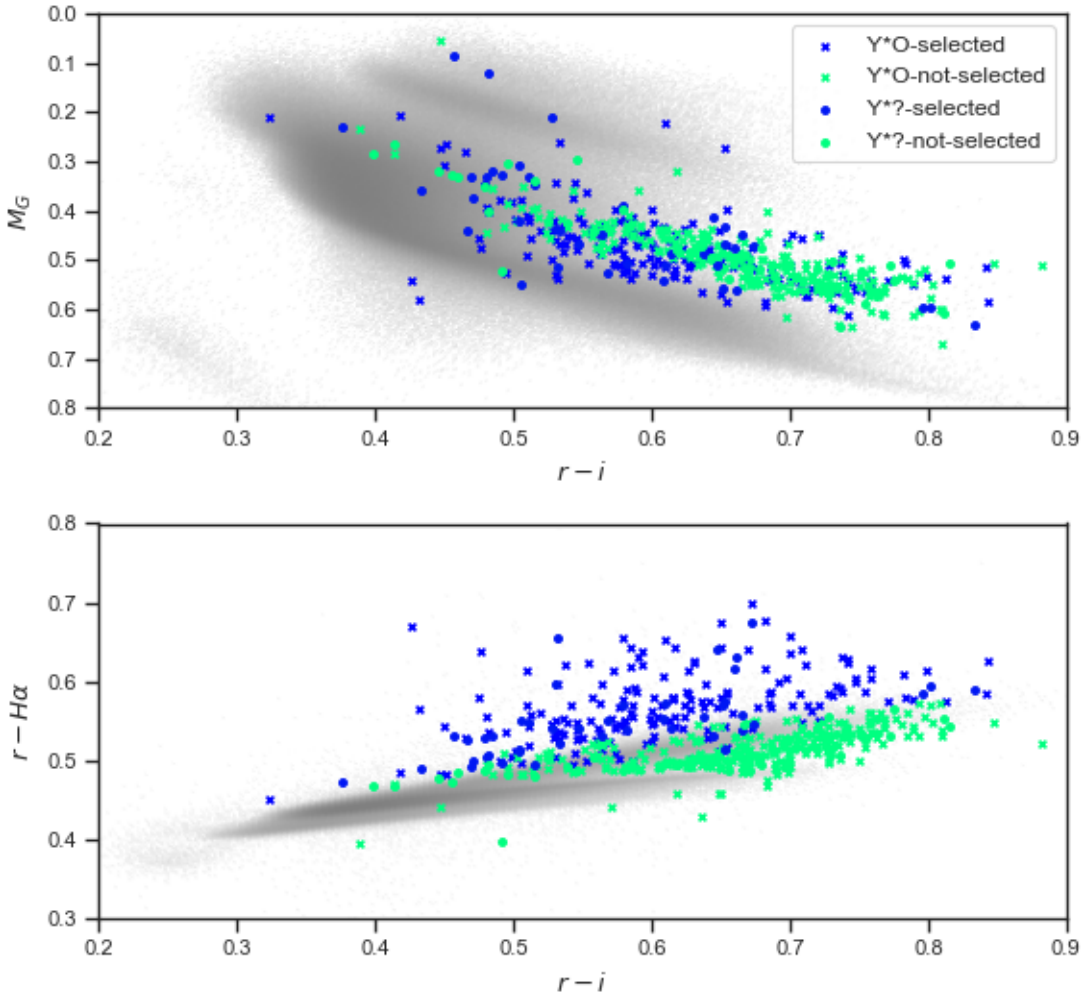


Figure 5.7: Young stellar and young stellar candidate objects from SIMBAD, with the ones selected in blue and the others in green.

sources that were classified as emitters by Witham et al. (2008) [66] have shifted from being quite clear emitters to positions on the upper locus in the colour-colour plot. Furthermore, 180 of these sources were classified as cluster sources by DBSCAN, indicating that, based on the Gaia/IPHAS catalogue, these sources are almost certainly not $H\alpha$ emitters. The remaining 176 "missed" emitters were manually investigated using the same plots as described in section 5.1. Based on those plots, most "missed" emitters did not look like $H\alpha$ emitters, but there were some that could be considered potential emitters. As discovered with the SIMBAD validation, these sources were missed either due to conservative thresholds or two loci in the selection neighbourhood, as discussed in more detail in section 5.3.

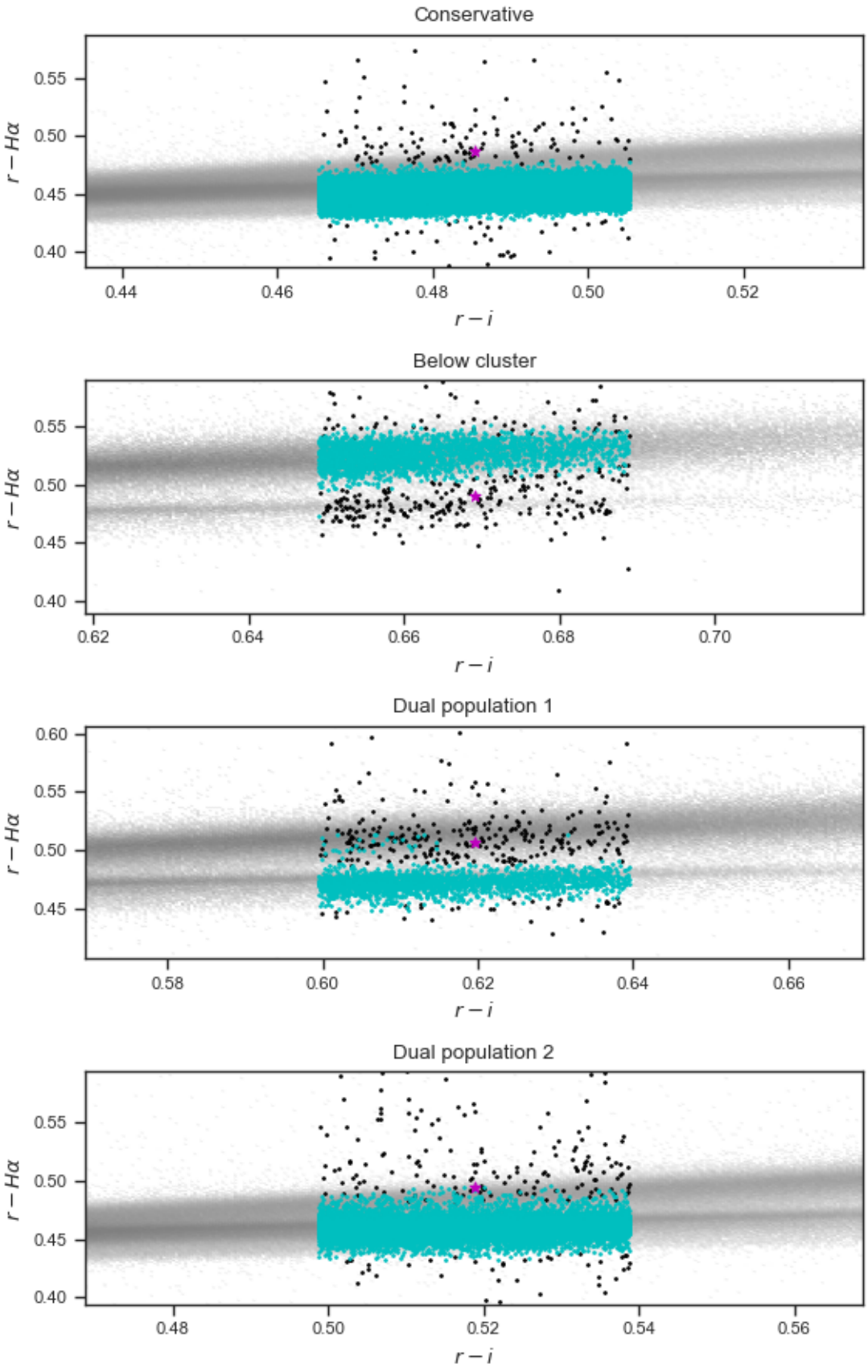


Figure 5.8: Different situations in which the selection algorithm failed to select emission line stars from SIMBAD. (*Top*) Example of SoI that was not classified as emitter due to the conservative thresholds used. (*Second from top*) SoI that is a SIMBAD emission line object, yet clearly sits below its locus. *Lower two panels* Both SoIs that failed to be identified as emitters due to their selection neighbourhood containing more than one locus.

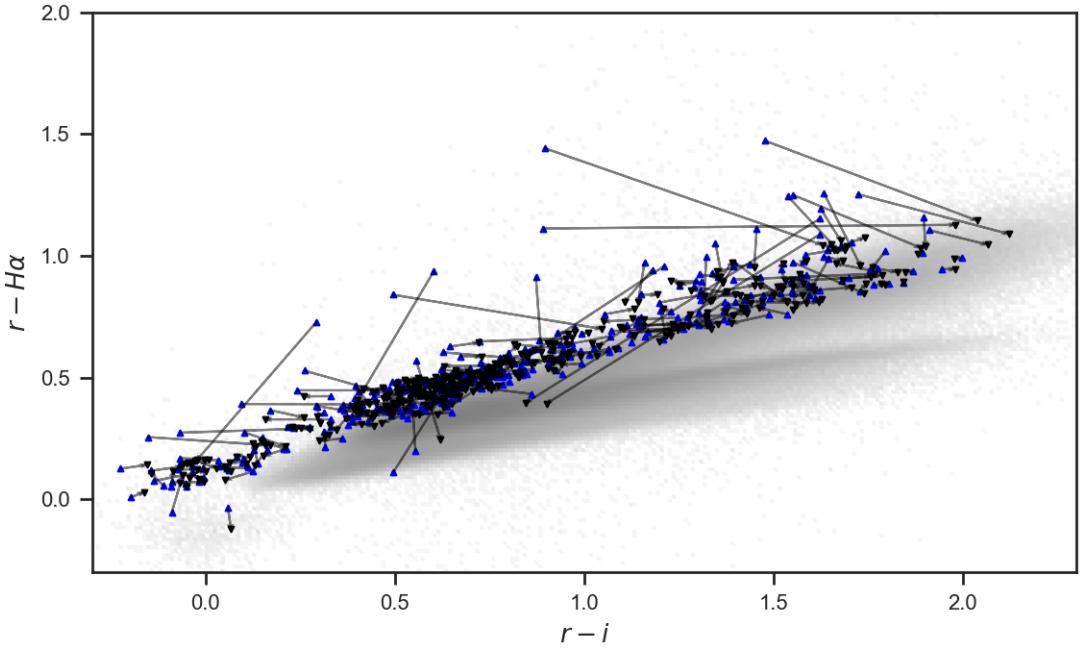


Figure 5.9: Colour-colour plot of the movement of the sources that were classified as emitters by Witham et al. (2008) [66] but not selected as $H\alpha$ emitters by the selection algorithm. Based on the Witham et al. (2008) [66] data some of these sources were quite clear $H\alpha$ emitters, but the majority of these have shifted onto the upper locus.

Chapter 6

Results and Discussion

In this chapter the results are discussed (section 6.2), the galactic longitude and latitude distribution of the emitters is investigated (section 6.1) and the limitations and possible improvements of the selection are discussed (section 6.3).

6.1 Distribution of Selected Emitters

In this section, the selected emitters are plotted along the galactic latitude and longitude, as shown in figure 6.1, and some of the regions of $H\alpha$ emitter over-densities are examined. To identify regions of over-densities the DBSCAN clustering algorithm was used in the l and b coordinate space, these clusters were then further investigated in the literature. A DBSCAN identified cluster of $H\alpha$ emitters (in l and b), is from here on referred to as group or $H\alpha$ emitter group, to prevent confusion with stellar clusters. The aim is to determine if the identified regions of $H\alpha$ emitter over-densities match up with regions known to harbour $H\alpha$ emitters and if this can be used to find potential new regions of interest.
The investigated regions of over-densities are listed below along with their respective approximate galactic longitude and latitude. Images from SDSS for most groups are also shown in the figures 6.2, 6.3, 6.4, 6.5 and 6.6 where the red circles highlight the position of selected emitters, with SIMBAD emission line objects shown as yellow crosses. Note: The images are in the IRCS coordinate system, not galactic longitude and latitude.

- **1.)** *l:* 202*, b:* 2.2*, figure 6.2* - These coordinates correspond to the NGC 2264 region, which contains the Cone Nebula and the Christmas Tree Cluster. The Cone Nebula is an HII region and the the Christmas Tree Cluster is a young open cluster. NGC 2264 belongs to the Monoceros OB1 association, a loose association of young main O and B type stars. It is well studied and known to contain a large number of $H\alpha$ emission objects such as T Tauri types [24, 41]. Figure 6.2 shows the NGC 2264 region with the positions of

identified $H\alpha$ emitters highlighted as red circles. The yellow crosses indicate the position of SIMBAD emission line objects.

- **2.)** *l:* 80*, b:* 2.6*, figure 6.3* - These emission line objects are part of the young open cluster NGC 6910, which is the core of the Cyg OB9 association and lies in the vicinity of the nebula IC 1318b, a part of IC 1318 – the Y Cygni Nebula. The NGC 6910 region is well studied. By the year 1990 more than 50 emission line objects were already known [31, 57]. It contains tens of Ae/Be Herbig and T Tauri stars and an additional 64 emitters were found in a region of 0.14 sq. deg. in 2011 [30].

- **3.)** *l:* 79*, b:* 0.35*, figure 6.3* - This region corresponds to the HII DR15 region on the southern periphery of the Cyg OB2 association, which lies in the star forming complex of the Cygnus-X region. The DR15 region has several confirmed $H\alpha$ emitters, with a majority of these likely to be young stellar objects [61, 37]. Out of the 11 selected emitters in this group, nine are classified as emission line objects by SIMBAD. The emitters' positions in this region are shown in figure 6.3 as red circles, with the yellow crosses indicating if it is also classified as emission line object in SIMBAD.

- **4.)** *l:* 84*, b:* 0.0*, figure 6.6* - This group of $H\alpha$ emitters lies inside the large W80 HII region, which contains the North America (NGC 7000) and Pelican (IC 5070) Nebulae, separated by the dust cloud that defines the "Atlantic Coast" and the "Gulf of Mexico". The North America (NGC 7000) and Pelican (IC 5070) Nebulae are known star formation regions, with some of these emitters already being recorded in the 1999 $H\alpha$ emitter survey by L. Kohoutek and R. Wehmeyer [26]. More recently more than 2000 young stellar objects were found in the North America and Pelican Nebulae using the Spitzer Space Telescope [23, 46]. The cluster of selected emitters at these coordinates lies in the northwest region (IRCS frame) of the Pelican Nebulae with their positions highlighted in figure 6.6 as red circles; with the yellow crosses indicating that this emitter is also classified as an emission line object in SIMBAD.

- **5.)** *l:* 85*, b:* -1.2*, figure 6.6* - These coordinates also lie inside the W80 HII region, sitting in the central dark region between the North America and Pelican Nebula, called the "Gulf of Mexico". The region is an active star formation region with at least 40 $H\alpha$ emitters and 35 Herbig-Haro objects [4]. The emitters in the "Gulf of Mexico" can be seen in figure 6.6.

- **6.)** *l:* 90*, b:* 2 - These coordinates lie in the L988 dark cloud complex close to an open cluster (C86) which contains about 50 $H\alpha$ emitters [25, 26]. Out of the ten selected

emitters, eight are emission line objects in SIMBAD with the majority of these further classified as potential or confirmed young stellar objects.

- **7.)** *l:* 99.5*, b:* 3.7*, figure 6.4* - These coordinates correspond to the IC 1396 HII region, shown in figure 6.4, which is part of the Cepheus OB2 association and has several distinctive features. In its centre lies the primary excitation source, HD 206267, and the open cluster Trumpler 37. To the west of this lies the Elephant Trunk Nebula (IC1396A). A large number of selected emitters are located near HD 206267 and the Elephant Trunk nebulae, shown in figure 6.4 as red circles. Both regions are known to harbour $H\alpha$ emitters and the Elephant Trunk Nebula is a known active star formation region [35, 34, 20]. North of HD 206267 lies the bright-rimmed cloud SFO 38, which is also a known star formation region with several tens of young stellar objects found in the vicinity, with a increased concentration of YSOs to the south of the bright rimmed cloud [16, 14]. The increased concentration of $H\alpha$ emitters can also be seen from the selected emitters shown as red circles in figure 6.4.

- **8.)** *l:* 109*, b:* 2.7*, figure 6.5* - This group of $H\alpha$ emitters lie in the active star formation region of the Cepheus OB3 association [32], to the east of the Cep B molecular clump as shown in figure 6.5. This region hosts a rich young open cluster, which is concentrated into two sub-clusters, one to the east of the Cep B molecular clump and the other near the Cep F molecular clump. Over 1000 young stellar objects have been found in this young cluster with the Spitzer Space Telescope [2]. Out of the 41 selected emitters in this region only one is in SIMBAD as an emission line object.

- **9.)** *l:* 109*, b:* 1.1 - The open cluster NGC 7419 lies at this position, hosting 31 Be stars, estimated to make up $36 \pm 7\%$ of all cluster B-type stars brighter than $R_C = 16.1$ mag [43, 28]; making NGC 7419 one of the Be star richest clusters. The group of selected emitters most likely belong to this cluster and out of the thirteen selected emitters in this region, six are classified as Be stars in SIMBAD.

- **10.)** *l:* 105*, b:* 3.8*, figure 6.7* - This region contains three groups of $H\alpha$ emitters, however only one (red) of these contains two SIMBAD sources; the others have no SIMBAD matches at all. Neither of the two SIMBAD matches is classified as an emission line object, their only reference is to the IPHAS paper[18]. The sources are highlighted in figure 6.7 and a histogram of their distances is shown in figure 6.8. The northern two groups, red and blue, sit in the dark cloud complex L 1188, which is a known active star formation region [1]; contains a spectroscopically confirmed young B-type star, thought to be a Herbig Be

star [67]. As both groups sit in a known active star formation region these could be two new young open clusters. It is however worth noting that the distances, shown in figure 6.8, vary quite largely for the blue group, which means that these probably do not belong to the same cluster. The third group of $H\alpha$ emitters sits further south near the reflection nebulae GN 22.14.9 and DG 182. Not much appears to be known to about this region, there is one SIMBAD emission line object about 2-3 arc-minutes to the east. These groups of $H\alpha$ sources can be followed up with Gaias' proper motions and distances, to determine if the sources in each group belong to the same cluster. Further spectroscopic follow up would allow additional classification of these sources, which could shed further light on their surrounding regions.

The fact that almost all further investigated regions of $H\alpha$ emitter over-densities are associated with regions known to contain $H\alpha$ emitting sources, such as young open clusters, star formations regions, OB associations, HII regions, is further confirmation that the selected sources selected are most likely $H\alpha$ emission line objects. Furthermore, it also showed that with some simple clustering, it was possible to find a new potential region of interest in terms of $H\alpha$ emitters.

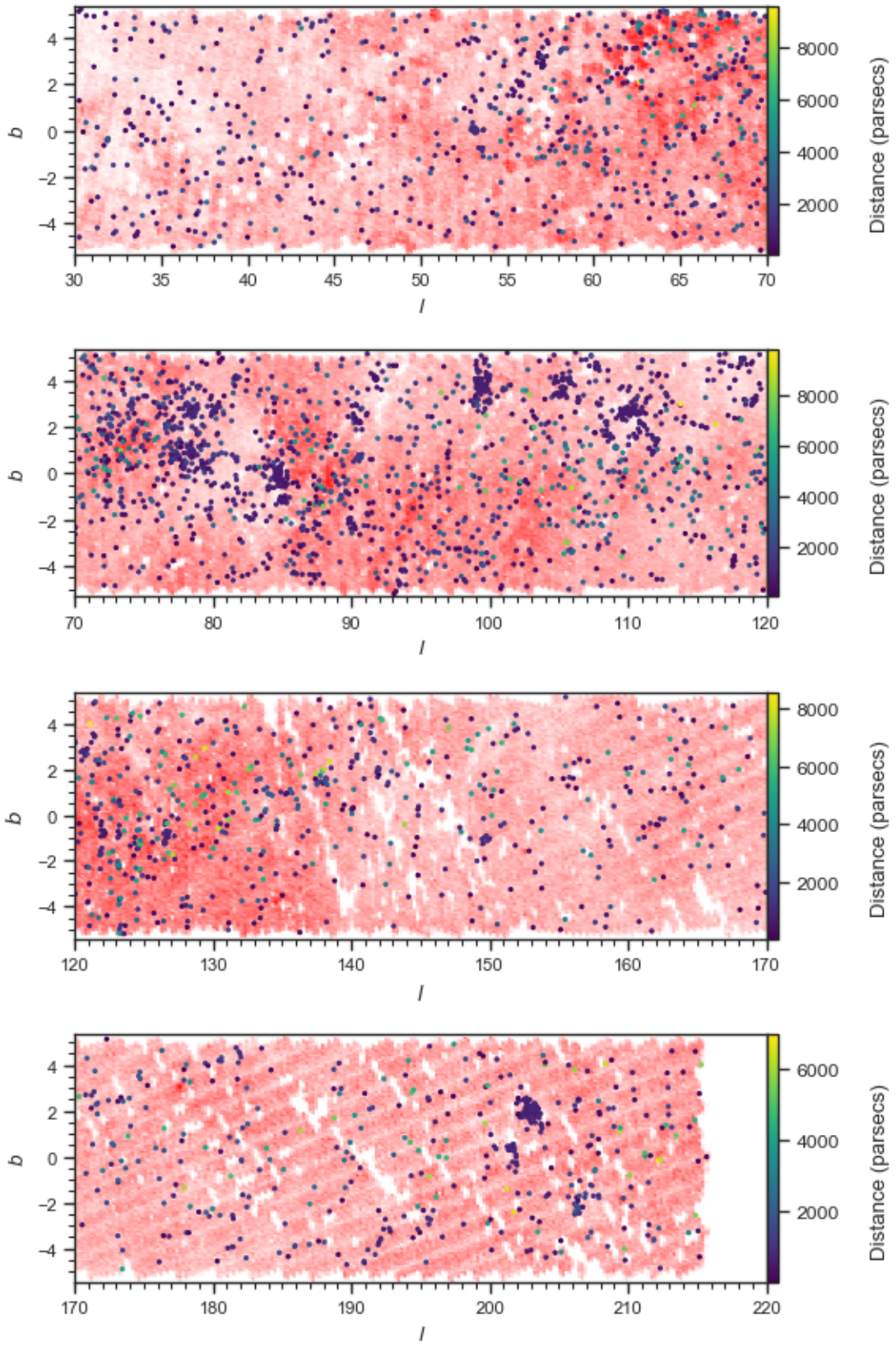


Figure 6.1: Galactic longitude and latitude distribution plot, with all sources from the Gaia/IHAS VAC catalogue shown as a density plot and with emitters shown as points. The distance of the emitters is indicated by the colouring. There are several region of $H\alpha$ over-densities; a brief description of some of these is given in section 6.1

Figure 6.2: Shown here is the NGC 2264 region, containing the distinctive Cone Nebula, an HII region, and the Christmas Tree Cluster, a young open cluster. NGC 2264 belongs to the well studied Monoceros OB1 association which contains a large number of confirmed $H\alpha$ emission objects. The position of selected $H\alpha$ emitters are shown as red circles, with SIMBAD emission line objects shown as yellow crosses. Contains group **1**.

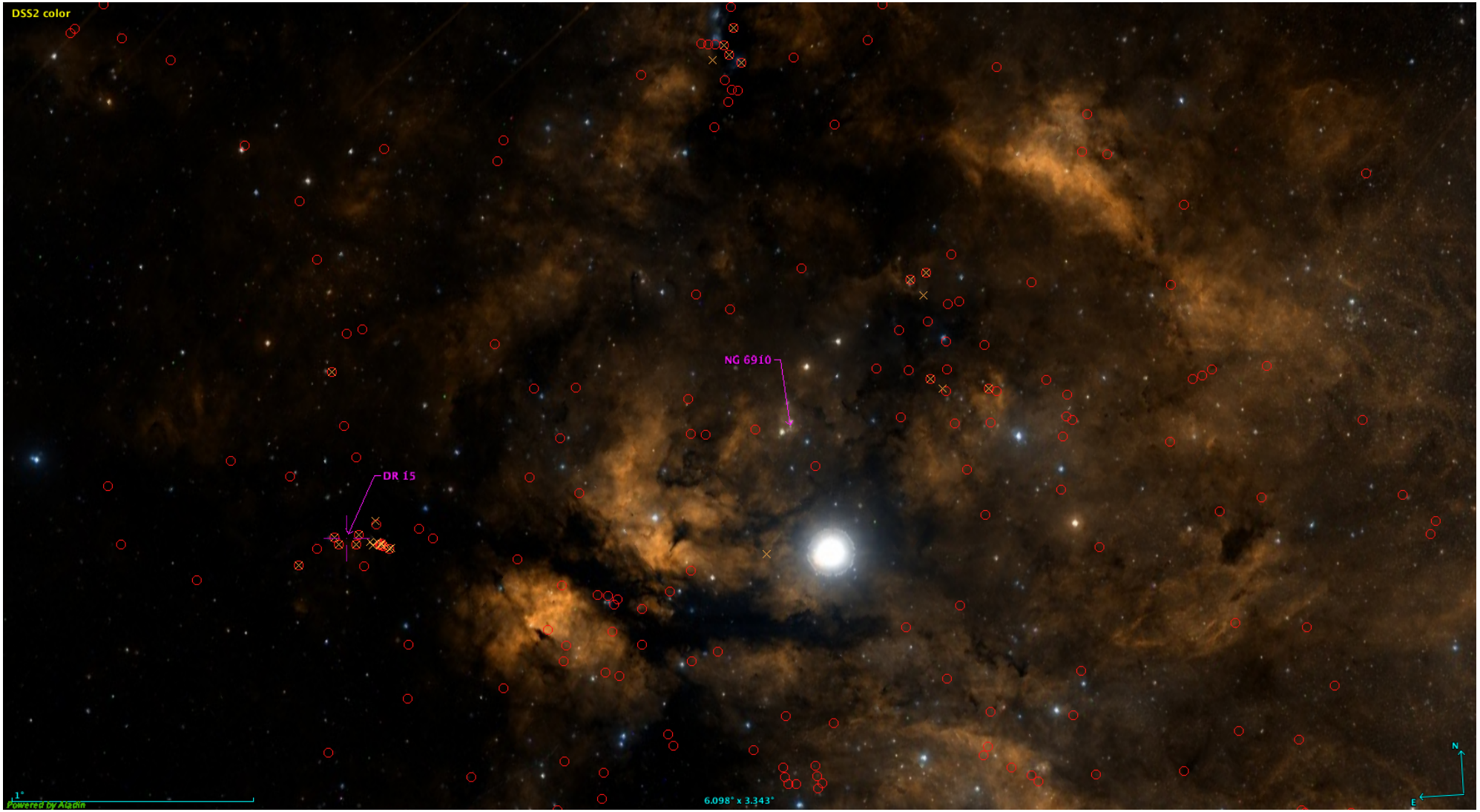


Figure 6.3: Shown here is the Cyg OB2 region with the young open cluster NGC 6910 at its core and the star Sadr (Gamma Cygni), one of the brightest visible stars. Cyg OB9 is part of the star forming region Cygnus-X region. The position of selected emitters are shown as red circles and the yellow crosses highlight the positions of the SIMBAD emission line objects. Of note is the group of emitters near the DR15 HII region, which has several confirmed $H\alpha$ emitters, with a majority of these likely young stellar objects. [61, 37] Contains groups **2** and **3**.

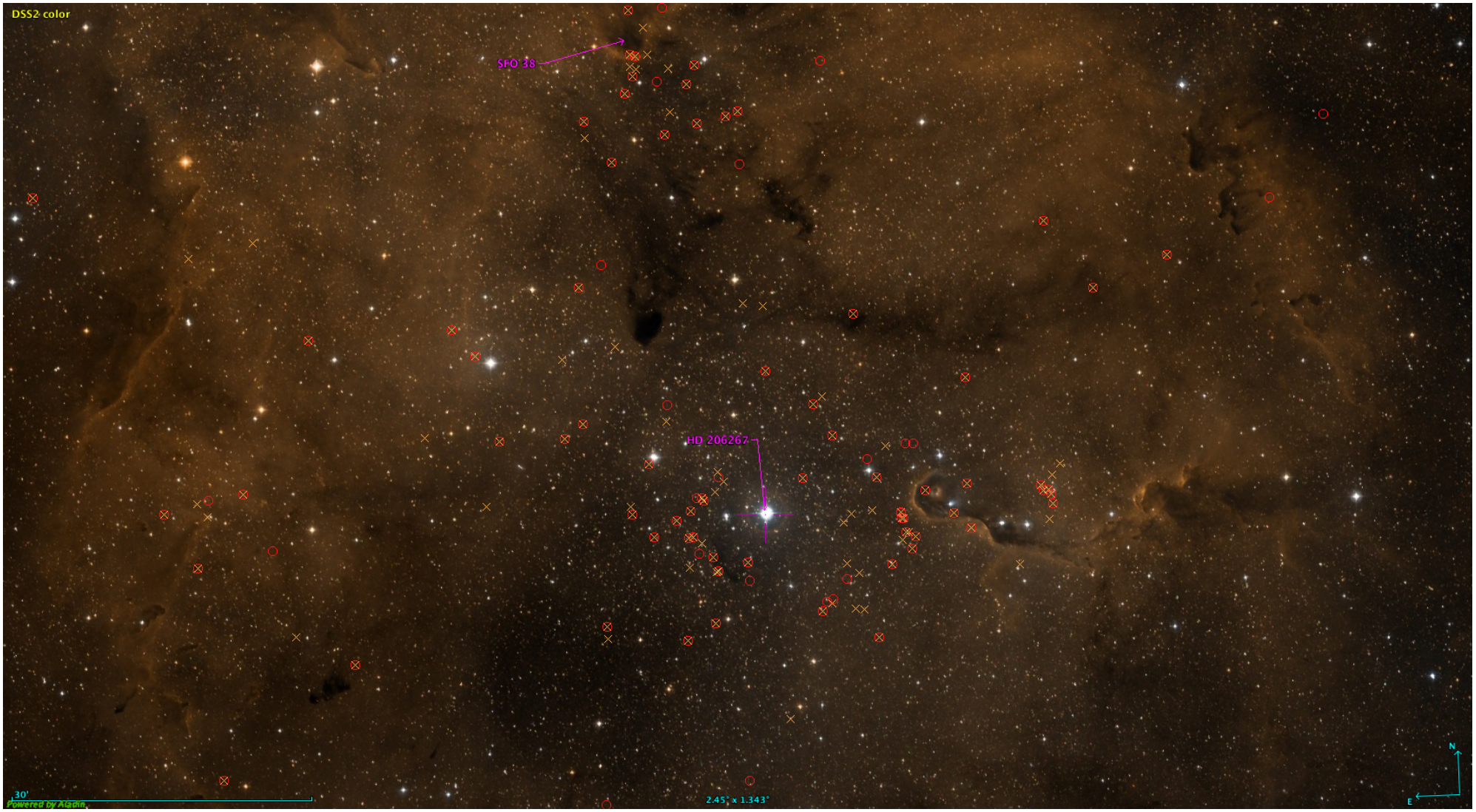


Figure 6.4: The IC 1396 HII region shown here, is part of the Cepheus OB2 association; with the primary excitation source, HD 206267 and the open cluster Trumpler 37 at its centre. The Elephant Trunk Nebula (IC 1396A) lies to the west of those, with both regions containing confirmed $H\alpha$ emitters and IC 1396A is known as an active star formation region [35, 34, 20]. In the north is the bright rimmed cloud SFO 38, which is also a star formation region several tens of young stellar objects found in and around SFO 38, with an increased concentration to the south [16, 14]. The red circles show the position of selected emitters and the yellow crosses show the emission line objects from SIMBAD. Contains group **2**

Figure 6.5: The group of selected emitters (red circles) shown here, lie to the east of the Cep B molecular clump in the active star formation region Cepheus OB3 association [32]. This region contains a large young open cluster with over 1000 young stellar objects found with the Spitzer space telescope [2]. Almost all of the selected $H\alpha$ emitters in this region are not classified as emission line objects in SIMBAD (yellow crosses), i.e. there is a good chance that some of these are not previously identified $H\alpha$ emitters. Contains group **8**.

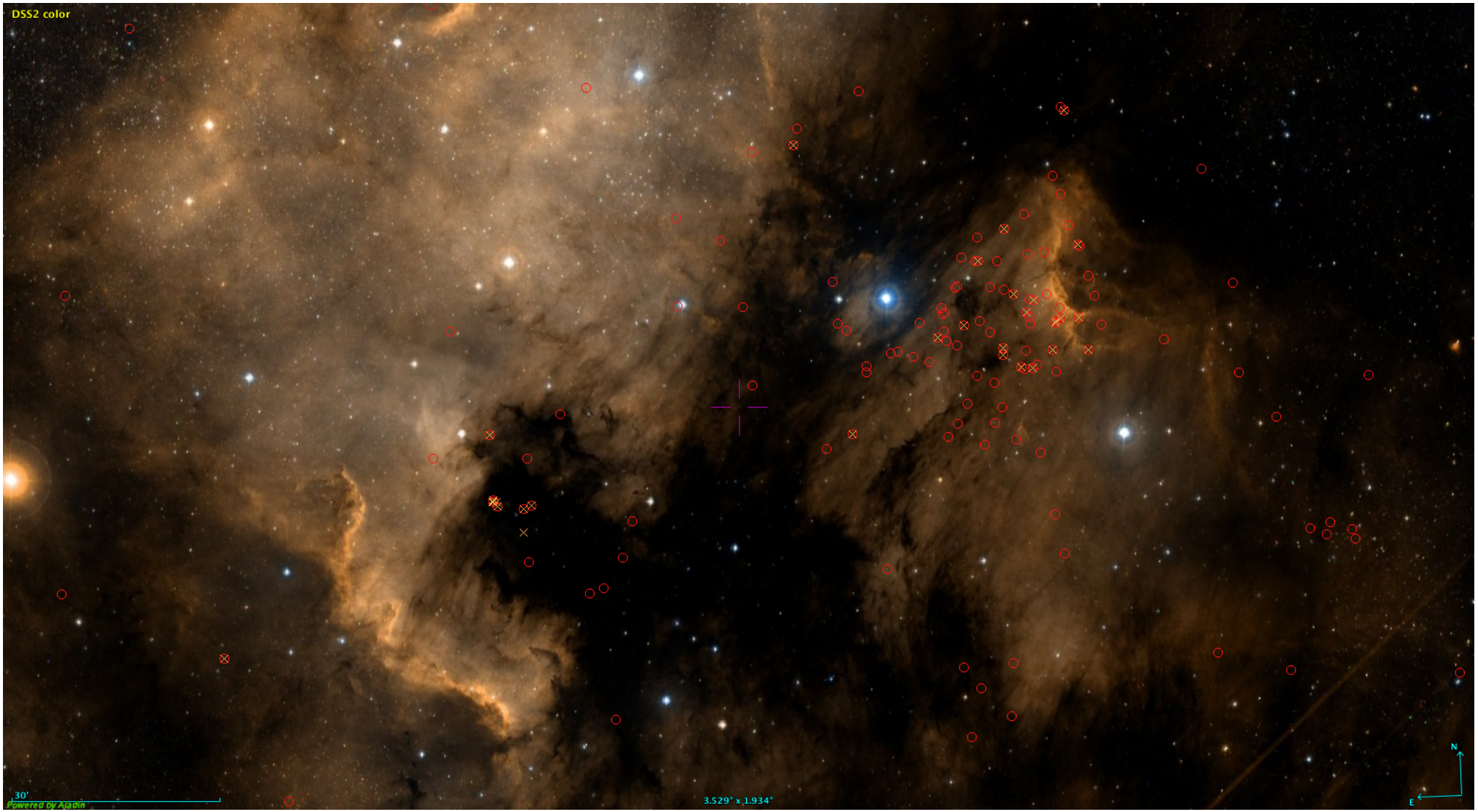


Figure 6.6: The W80 HII region shown here, contains the North America (NGC 7000) and Pelican (IC 5070) Nebula, separated by the dust cloud that defines the "Atlantic Coast" and the "Gulf of Mexico". With more than 2000 identified young stellar objects found across the two nebulae, both are known star formation regions [23, 46]. From the selected emitters the group in the northwest region of the Pelican nebula and the "Gulf of Mexico" are noteworthy. Contains groups **4** and **5**.

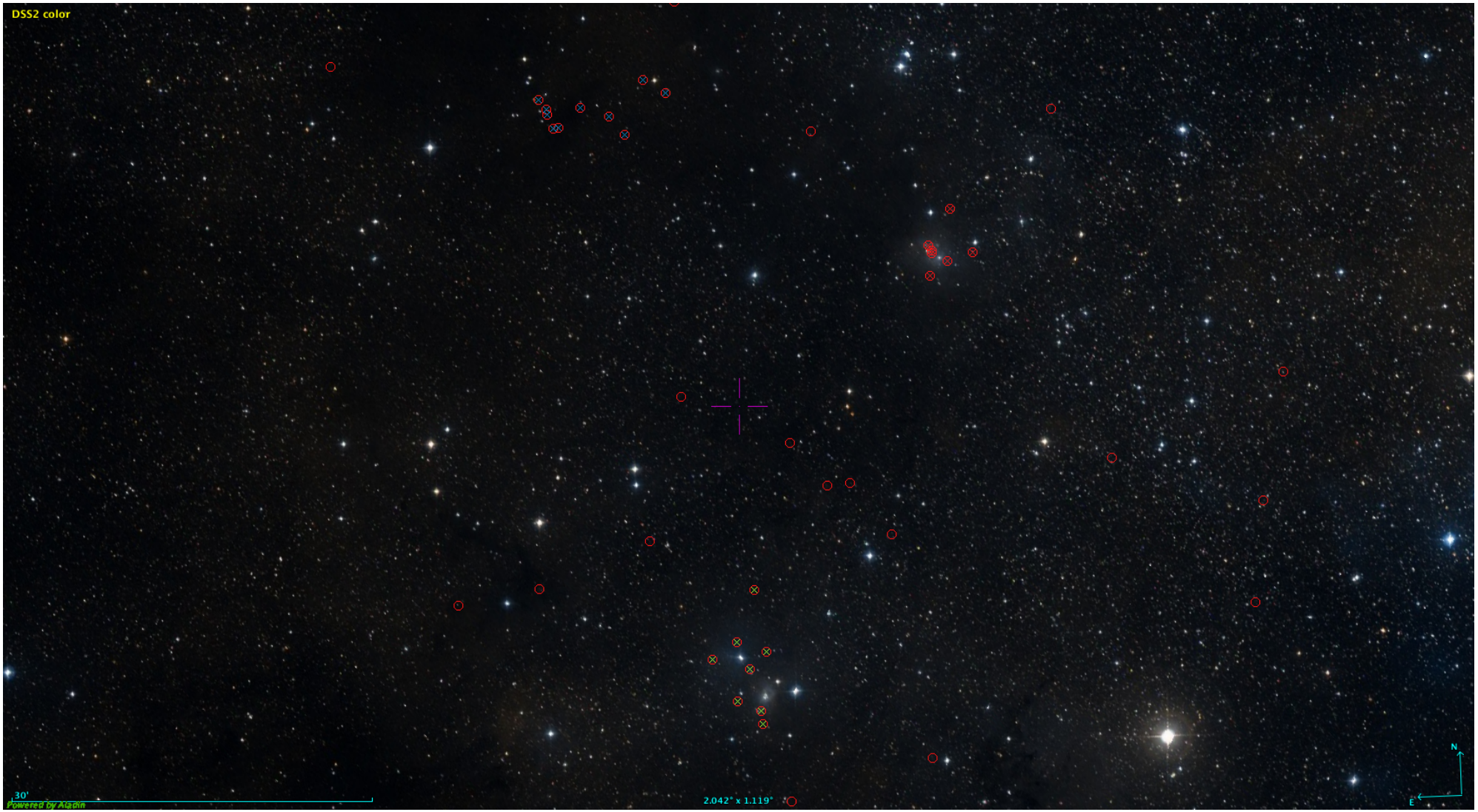


Figure 6.7: Region containing three groups of $H\alpha$ emitters, with only two the sources known to SIMBAD and neither classified as $H\alpha$ emitter. The northern two groups lie in the dark cloud complex L1118, an active star formation region [1, 67]. The group further south is located close to the two reflection nebulae GN 22.14.9 and DG 182; based on SIMBAD, not much is known about this region. There is one SIMBAD object, classified as emission line object, to the east.

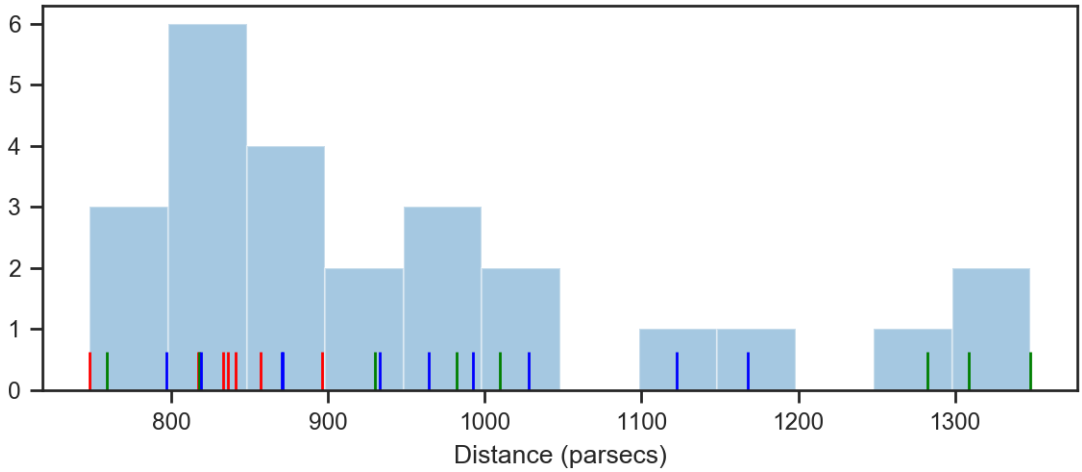


Figure 6.8: The distance distribution of the sources shown in figure 6.7, with the colour indicating to which group the source belongs. The red group is clustered reasonably tightly, whereas the green and blue group have a much larger spread.

6.2 Results

The work described above resulted in a new dataset of 21,381 sources with 3,447 selected as excess $H\alpha$ emission line sources based on the thresholds described in section 4.2. The dataset contains the full 21,381 SoIs as the scores (EP, WEP, NCN and kNCN) were calculated for all SoIs.

The selected emitters, based on the thresholds used, are shown in figures 6.9 and 6.10. This catalogue provides a valuable resource for finding excess $H\alpha$ sources, with respect to their population, which can then be further investigated spectroscopically. Many of the selected $H\alpha$ emitters, such as the one in figure 5.3, are most likely newly found $H\alpha$ emitters, as only 1,183 of the 3,447 selected sources are either in SIMBAD or Witham et al. (2008) [66]. Furthermore, combined with Gaia's proper motion and distance observations, it can be used for finding potential new young open clusters.

As discussed in section 1.3 in the introduction, one of the aims was to find reddened $H\alpha$ emitters, that can't be found by purely using a colour-colour plot. Figure 6.10 shows that the selection algorithm was able to do this to a certain degree. For lower r-i values, there are many selected emitters that sit in the upper locus, however towards large values of r-i the number of selected sources just above the lower locus drops significantly, which at least to some degree is due to the problem of two populations in a selection neighbourhood of a SoI, as discussed in section 5.3.

As already mentioned in section 5.3, the set of selected $H\alpha$ emitters has, based on the validation performed (chapter 5), a low number of incorrectly selected sources. This is due to the fact that

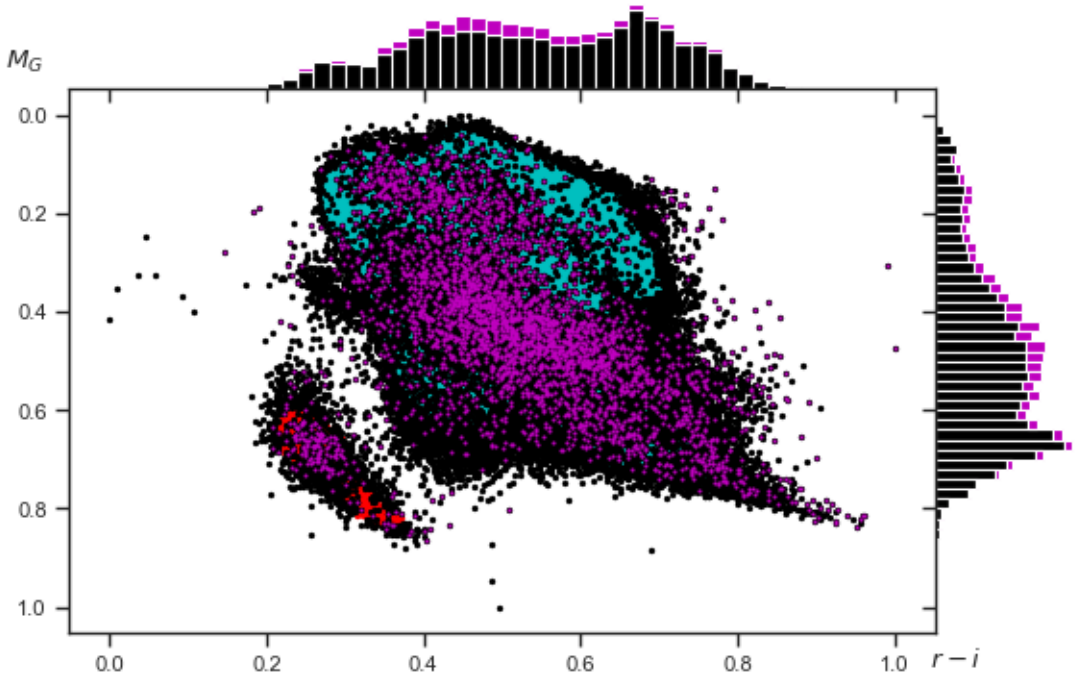


Figure 6.9: Colour-magnitude plot of selected emitters (purple) with "noise" sources shown in black.

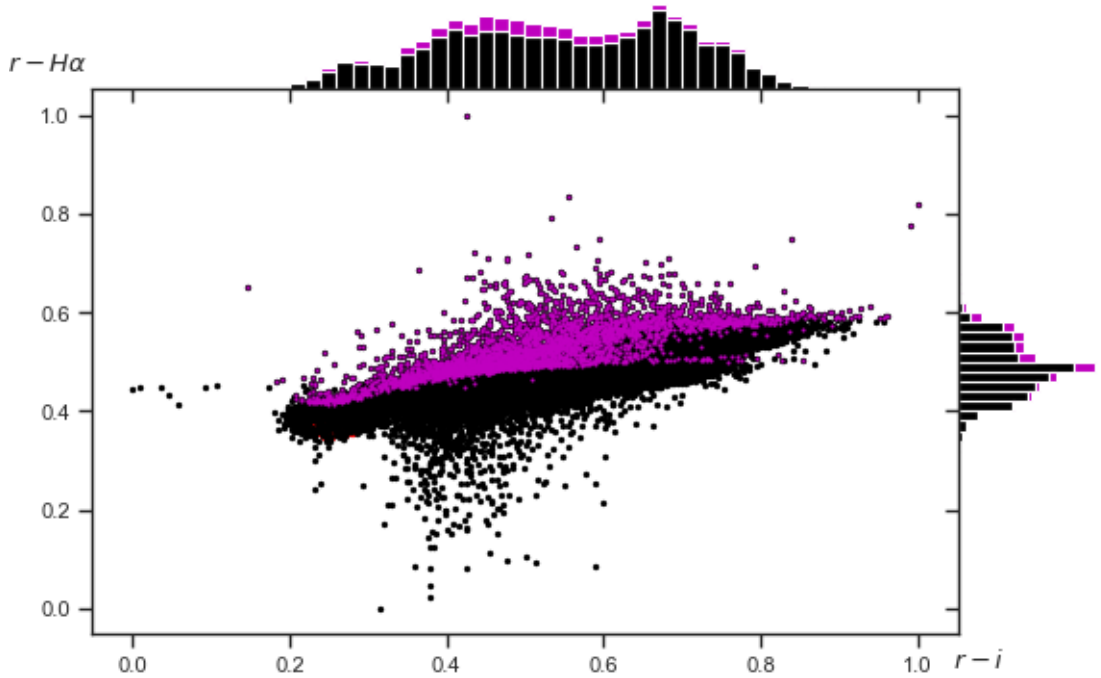


Figure 6.10: Colour-colour plot of the emitters (purple) and "noise" sources shown in black. Noteworthy are the selected emitters that are sitting on the unreddened main-sequence branch, these are sources that would been missed without the additional Gaia information as discussed in section 1.3.

the cleanliness of the set of selected emitters was deemed more important than completeness, and the score thresholds were chosen accordingly. However, as the resulting catalogue includes

all SoIs, i.e. all sources classified as "noise" by DBSCAN, and their associated scores, re-selection of $H\alpha$ emitters using different thresholds for a more aggressive selection approach is straightforward. To further improve the completeness of the sample, would also require addressing some of the limitations of the selection algorithm, as discussed in section 6.3.

6.3 Limitations, Difficulties and Improvements of the Selection Algorithm

Most of the limitations and difficulties of the selection algorithm have already been briefly mentioned in some of the sections above, the following is a summary:

- **DBSCAN parameter selections**: Choosing the correct DBSCAN parameters is one of the more difficult steps for this selection algorithm as there is no easy way of identifying the correct parameters and neither of the two parameters (ϵ and $MinPts$) is very intuitive. While the clustering algorithm is able to identify clusters of any shape, it is limited, when identifying clusters of vastly different densities, as the cluster density is set via the parameters. Additionally, the gridding flattens the population density differences significantly, however two seperate DBSCAN runs, with different parameters, were still required to identify the main sequence and white dwarf cluster sufficiently. This adds unnecessary complexity, and a possible future improvement would be to try and use a different clustering algorithm instead, such as OPTICS, which is a generalisation of DBSCAN that does not require the ϵ parameter and produces hierarchical results related to that of linkage clustering.

- **Dual population**: The selection algorithm proposed is not able to select emitters that contain two loci in their selection neighbourhood as it will always use the distance to the closest cluster points in the r-$H\alpha$ to calculate the kNCN score; the last two panels in figure 5.8 are an example of this scenario. The SIMBAD emission line objects, most likely belong to the lower branch. However, the selection algorithm is unable to take this into account and the cluster sources from the unreddened main sequence are used to calculate the kNCN score, which means that these type of emitters were consequently missed by the selection algorithm as their kNCN scores don't meet the required threshold. It is worth noting that this scenario only occurs when cluster sources from both loci are in the selection neighbourhood, if it only contains a single loci then these type of emitters would be selected. Therefore one approach that might prevent this problem, is to reduce the

tunnelling rectangle on the CMD plane, for sources in dense regions.

- **Slicing** Identifying of the relevant locus to use for the selection neighbourhood is only done in single directions in the CMD. This could be improved by using an extending radius that keeps growing until the closest cluster contains enough data sources in the neighbourhood. This modification would also remove the need for manual selection of reference points for the few SoIs for which slicing was unsuitable. Another limitation of slicing is that SoIs will always be classified with respect to the sources on the edge of the population, however, this would not be fixed by the modified approach proposed (unless the number of required cluster sources would be increased appropriately).

Chapter 7

Conclusion

Many interesting types of stars show excess $H\alpha$ emission, making identification of $H\alpha$ emitters an important step for increasing the sample sizes of these type of stars and allow further improvement of our models and understanding. Pre- and post-main sequence stars, interacting binaries and classical Be stars are some of the types of stars that exhibiting strong $H\alpha$ emission; identifying $H\alpha$ emitting sources from photometric surveys such as IPHAS is helpful for target selection for spectroscopic follow-up observations. Furthermore combining an $H\alpha$ survey with a photometric survey dataset that includes the absolute magnitudes, such as Gaia, offers further improvement, as sources of interest can be placed accurately on a CMD giving an idea of the type of object, since different types of stars, such as YSO and Be stars, are found in different regions of the CMD.

Presented here is a catalogue of 3,447 selected $H\alpha$ emitters, based on the Gaia/IPHAS VAC catalogue produced by Simone Scaringi et al. (2018) [51]. The emitter selection process uses unsupervised machine learning techniques, hence a brief overview is given, explaining its potential for analysis and classification of astronomical objects. The density-based clustering algorithm used is DBSCAN, capable of identifying clusters of arbitrary shape with a constant density defined by two parameters: ϵ, a radius of sorts, and $MinPts$. It is used in the three dimensional colour-magnitude-colour space of r-i, M_G and r-$H\alpha$ to identify the main-sequence and giant cluster, and white dwarf cluster. Unlike many other clustering algorithms, such as k-means, DBSCAN does not assign a cluster to every data point; instead sources that are not classified as part of a cluster are labelled as "noise". This is used to identify the local outliers in the three dimensional space, which are then treated as potential emitters. Clearly not all of these are actually outliers in the $H\alpha$ dimensions with respect to their local population, many of them will just be an a outlier in the CMD plane; so the local neighbourhood is used in the r-$H\alpha$ dimension to calculate two scores: the weighted empirical probability (WEP) and the k-nearest cluster neighbours (kNCN) distance to determine a source of interest's (SoI) r-$H\alpha$ deviation

from the main locus in their local neighbourhood.

In order for the scoring to be done with respect to the sources population, the local neighbourhood has to include the relevant population; for sources above the cluster this is achieved by selecting all sources within a rectangular area around the SoI in the CMD plane, i.e. ignoring r-$H\alpha$ values as this is the dimension of interest. However, for sources that are beside a cluster in the colour-magnitude plane, this is not as straightforward. Therefore slices of a narrow width in the colour-magnitude plane in the r-i and M_G dimension are used to identify the nearest cluster and include it in the "local neighbourhood", which is then used to calculate the scores. One downside of this approach is that these SoIs are always classified with respect to the r-$H\alpha$ values of the edge of the population.

Once the selection neighbourhood (the sources used to score a SoI) have been determined for all SoIs, and the WEP and kNCN scores calculated, emitters are selected using thresholds of the scores. The WEP threshold is constant and requires all emitters to be above 0.92; in other words, have r-$H\alpha$ values in the top 92% of their respective selection neighbourhood. The thresholds for the kNCN score are an exponential function of WEP, i.e. as WEP decreases, the kNCN score has to increase in order for a SoI to be selected as an $H\alpha$ emitter.

In order to check the accuracy of the selected $H\alpha$ emitters, these were split into six groups using the two scores, with the groups decreasing in their r-$H\alpha$ deviation from their locus (based on the scores). Each group was then crossmatched with the LAMOST dataset, which contains spectral data. The spectra of the matched sources were then manually validated to confirm if these are $H\alpha$ emitters or not. As expected, at lower scores the number of incorrect $H\alpha$ emitters increases slightly. However, overall the accuracy based on LAMOST comparison is reasonably high; out of the 96 spectra manually viewed, only nine were incorrectly classified as emitters. Further comparison was also done with the Witham et al. (2008) emitter catalogue [66] and the SIMBAD dataset. Cross-matching with SIMBAD was simple as it supports the Gaia DR2 identifier. All SoIs were therefore crossmatched with SIMBAD, giving 1,437 matches with 402 of these classified as emission line star type by SIMBAD. Out of these 402 emitters, 270 were selected by the selection algorithm covered here. Investigation of the emitters missed showed that the main reasons for not selecting these emitters were: a) the thresholds are set quite conservative to produce a clean set $H\alpha$ emitters, with a minimal number of incorrect selections and b) emitters that belonged to the reddened main-sequence and giant branch in the colour-colour plot with their selection neighbourhood containing sources from the unreddened branch, which results in a low kNCN score and therefore not meeting the required selection thresholds. Overall the validation showed that the selection algorithm works well in identifying $H\alpha$ emitters. However, completeness is limited due to conservative thresholds and limitations in the selection

algorithm.

The selected emitters were also plotted in the galactic coordinates to see if they matched up with any regions known to harbour $H\alpha$ emitters, such as star formation regions, HII regions etc. To achieve this DBSCAN clustering was applied to the selected emitters in galactic longitude and latitude space. The resulting clusters were then investigated and checked against the literature. Many of these clusters of emitters matched up with regions that are expected to have an increased density of $H\alpha$ emitters, such as OB associations, HII nebulae and young open clusters. Three new (i.e. no references found in Simbad) groups of $H\alpha$ emitters were also found, with at least one of these being a potential new young open cluster based on the emitters distance distribution.

Overall this approach showed that unsupervised machine learning can be successfully used to assist in selection of sources of interest in a given feature space; the value of combining $H\alpha$ surveys with other photometric surveys that include the distance and therefore absolute magnitude. Validation showed that this sample only contains a small number of incorrectly-selected emitters and that many of these selected $H\alpha$ emitters are possibly new, as only 1,183 of the 3,447 selected sources are found in either SIMBAD or Witham et al. (2008) [66]. However, due to conservative selection thresholds and limitations with the selection algorithm for reddened emitter sources that have unreddened sources in their selection neighbourhood, the sample's completeness is limited. In terms of future work, spectroscopically follow-up observations would provide further information on the performance of the selection algorithm. Furthermore addressing some of the limitations of this selection algorithm should allow for improvement in the completeness of the dataset of $H\alpha$ emitters.

Bibliography

[1] Peter Abraham, Kazuhito Dobashi, Akira Mizuno, and Yasuo Fukui. "Molecular material and young stellar objects in the L 1188 dark cloud complex." In: *Astronomy and Astrophysics* 300 (1995), p. 525.

[2] Thomas S Allen, Robert A Gutermuth, Erin Kryukova, S Thomas Megeath, Judith L Pipher, Tim Naylor, RD Jeffries, Scott J Wolk, Brad Spitzbart, and James Muzerolle. "Spitzer Imaging of the Nearby Rich Young Cluster, Cep OB3b". In: *The Astrophysical Journal* 750.2 (2012), p. 125.

[3] Domenica Arlia and Massimo Coppola. "Experiments in parallel clustering with DBSCAN". In: *European Conference on Parallel Processing*. Springer. 2001, pp. 326–331.

[4] Tina Armond, Bo Reipurth, John Bally, and Colin Aspin. "Star formation in the "Gulf of Mexico"". In: *Astronomy & Astrophysics* 528 (2011), A125.

[5] Yu Bai, Ji-Feng Liu, and Song Wang. "Machine learning classification of Gaia Data Release 2". In: *Research in Astronomy and Astrophysics* 18.10 (2018), p. 118.

[6] Geert Barentsen, Hywel John Farnhill, JE Drew, EA González-Solares, R Greimel, MJ Irwin, Brent Miszalski, Christine Ruhland, P Groot, A Mampaso, et al. "The second data release of the INT Photometric Hα Survey of the Northern Galactic Plane (IPHAS DR2)". In: *Monthly Notices of the Royal Astronomical Society* 444.4 (2014), pp. 3230–3257.

[7] Charles Beichman, Bjoern Benneke, Heather Knutson, Roger Smith, Pierre-Olivier Lagage, Courtney Dressing, David Latham, Jonathan Lunine, Stephan Birkmann, Pierre Ferruit, et al. "Observations of transiting exoplanets with the James Webb Space Telescope (JWST)". In: *Publications of the Astronomical Society of the Pacific* 126.946 (2014), p. 1134.

[8] Jon Louis Bentley. "Multidimensional binary search trees used for associative searching". In: *Communications of the ACM* 18.9 (1975), pp. 509–517.

[9] AGA Brown, A Vallenari, T Prusti, JHJ de Bruijne, C Babusiaux, CAL Bailer-Jones, Gaia Collaboration, et al. "Gaia Data Release 2. Summary of the contents and survey properties". In: *arXiv preprint arXiv:1804.09365* (2018).

[10] *By Antti Ajanki - Own work.* URL: `https://commons.wikimedia.org/wiki/File:KnnClassification.svg`.

[11] *By Chire - Own work, CC BY-SA 3.0.* URL: `https://commons.wikimedia.org/w/index.php?curid=17085332`.

[12] Bradley W Carroll and Dale A Ostlie. *An introduction to modern astrophysics*. Cambridge University Press, 2007.

[13] Olivier Chapelle. *Semi-Supervised Learning*. ProQuest Ebook Central: MIT Press, 2006. ISBN: 9780262255899.

[14] Neelam Chauhan, AK Pandey, K Ogura, DK Ojha, BC Bhatt, SK Ghosh, and PS Rawat. "Triggered star formation and evolution of T-Tauri stars in and around bright-rimmed clouds". In: *Monthly Notices of the Royal Astronomical Society* 396.2 (2009), pp. 964–983.

[15] Arnab Rai Choudhuri. *Astrophysics for physicists*. Cambridge University Press, 2010.

[16] Rumpa Choudhury, Bhaswati Mookerjea, and HC Bhatt. "Triggered star formation and young stellar population in bright-rimmed cloud SFO 38". In: *The Astrophysical Journal* 717.2 (2010), p. 1067.

[17] Dinko P Dimitrov, Diana P Kjurkchieva, and Emil I Ivanov. "A Study of the Hα Variability of Be Stars". In: *The Astronomical Journal* 156.2 (2018), p. 61.

[18] Janet E Drew, Robert Greimel, Mike J Irwin, Amornrat Aungwerojwit, Michael J Barlow, Romano LM Corradi, Jeremy J Drake, Boris T Gänsicke, P Groot, A Hales, et al. "The INT photometric Hα survey of the northern Galactic plane (IPHAS)". In: *Monthly Notices of the Royal Astronomical Society* 362.3 (2005), pp. 753–776.

[19] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231.

[20] Konstantin V Getman, Eric D Feigelson, Aurora Sicilia-Aguilar, Patrick S Broos, Michael A Kuhn, and Gordon P Garmire. "The Elephant Trunk Nebula and the Trumpler 37 cluster: contribution of triggered star formation to the total population of an HII region". In: *Monthly Notices of the Royal Astronomical Society* 426.4 (2012), pp. 2917–2943.

[21] Eduardo A González-Solares, Nicholas A Walton, Robert Greimel, Janet E Drew, Mike J Irwin, Stuart E Sale, K Andrews, Amornrat Aungwerojwit, Michael J Barlow, E Van Den Besselaar, et al. "Initial data release from the INT Photometric Hα survey of the northern Galactic plane (IPHAS)". In: *Monthly Notices of the Royal Astronomical Society* 388.1 (2008), pp. 89–104.

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 2672–2680. URL: `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

[23] S Guieu, LM Rebull, JR Stauffer, LA Hillenbrand, JM Carpenter, A Noriega-Crespo, DL Padgett, DM Cole, SJ Carey, KR Stapelfeldt, et al. "The North American and Pelican Nebulae. I. IRAC Observations". In: *The Astrophysical Journal* 697.1 (2009), p. 787.

[24] George H Herbig. “Emission-Line Stars Associated with the Nebulous Cluster NGC 2264.” In: *The Astrophysical Journal* 119 (1954), p. 483.

[25] GH Herbig and SE Dahm. “The Pre-Main-Sequence Population of L988”. In: *The Astronomical Journal* 131.3 (2006), p. 1530.

[26] L Kohoutek and R Wehmeyer. “Catalogue of H-alpha emission stars in the Northern Milky Way”. In: *Astronomy and Astrophysics Supplement Series* 134.2 (1999), pp. 255–256.

[27] L Lindegren, J Hernández, A Bombrun, S Klioner, U Bastian, M Ramos-Lerate, A De Torres, H Steidelmüller, C Stephenson, D Hobbs, et al. “Gaia Data Release 2-The astrometric solution”. In: *Astronomy & Astrophysics* 616 (2018), A2.

[28] SL Malchenko and AE Tarasov. “B and Be-stars in the young open stellar clusters NGC 659 and NGC 7419”. In: *Astrophysics* 54.1 (2011), p. 52.

[29] Dan Maoz. *Astrophysics in a Nutshell*. Princeton University Press, 2016.

[30] N. D. Melikian, V. S. Tamazian, A. A. Karapetian, A. L. Samsonian, and G. R. Kostandian. “New Hα stars. NGC 6910 region. II.” In: *Astrophysics* 54.2 (June 2011), pp. 203–213. ISSN: 1573-8191. DOI: 10.1007/s10511-011-9172-y. URL: https://doi.org/10.1007/s10511-011-9172-y.

[31] ND Melikian and VS Shevchenko. “Halpha Emission Stars in the NGC6910 Region”. In: *Astrofizika* 32 (1990), p. 169.

[32] Takao Mikami and Katsuo Ogura. “Hα Emission Stars in the Cepheus OB3 Region”. In: *Astrophysics and Space Science* 275.4 (2001), pp. 441–462.

[33] Tom M. Mitchell. *Machine learning*. McGraw Hill series in computer science. McGraw-Hill, 1997. ISBN: 978-0-07-042807-2. URL: http://www.worldcat.org/oclc/61321007.

[34] M Morales-Calderón, JR Stauffer, L Rebull, BA Whitney, D Barrado y Navascués, DR Ardila, I Song, TY Brooke, L Hartmann, and N Calvet. “Mid-Infrared Variability of protostars in IC 1396A”. In: *The Astrophysical Journal* 702.2 (2009), p. 1507.

[35] M Nakano, K Sugitani, M Watanabe, N Fukuda, D Ishihara, and M Ueno. “Wide-field Survey of Emission-line Stars in IC 1396”. In: *The Astronomical Journal* 143.3 (2012), p. 61.

[36] Elisabeth R Newton, Jonathan Irwin, David Charbonneau, Perry Berlind, Michael L Calkins, and Jessica Mink. “The Hα emission of nearby M dwarfs and its relation to stellar rotation”. In: *The Astrophysical Journal* 834.1 (2017), p. 85.

[37] EH Nikoghosyan, T Yu Magakian, and TA Movsessian. “Searches for HH-objects and emission stars in star-formation regions. VIII. Stars with Hα emission in the vicinity of the nebula GM 2-41”. In: *Astrophysics* 55.1 (2012), pp. 70–80.

[38] Andrei Novikov. *annoviko/pyclustering: pyclustering 0.8.2 release*. Nov. 2018. DOI: 10.5281/zenodo.1491324. URL: https://doi.org/10.5281/zenodo.1491324.

[39] M Ntampaka, J ZuHone, D Eisenstein, D Nagai, A Vikhlinin, L Hernquist, F Marinacci, D Nelson, R Pakmor, A Pillepich, et al. "A Deep Learning Approach to Galaxy Cluster X-ray Masses". In: *arXiv preprint arXiv:1810.07703* (2018).

[40] Harry Nyquist. "Certain topics in telegraph transmission theory". In: *Transactions of the American Institute of Electrical Engineers* 47.2 (1928), pp. 617–644.

[41] K. Ogura. "New H-alpha-emission stars in Monoceros OB1 and R1 associations". In: *pasj* 36 (1984), pp. 139–148.

[42] Mostofa Ali Patwary, Diana Palsetia, Ankit Agrawal, Wei-keng Liao, Fredrik Manne, and Alok Choudhary. "A new scalable parallel DBSCAN algorithm using the disjoint-set data structure". In: *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. IEEE Computer Society Press. 2012, p. 62.

[43] A Pigulski and G Kopacki. "NGC 7419: An open cluster rich in Be stars". In: *Astronomy and Astrophysics Supplement Series* 146.3 (2000), pp. 465–469.

[44] Onno Rudolf Pols. *Stellar structure and evolution*. Astronomical Institute Utrecht, 2011.

[45] Sebastian Raschka. *Python Machine Learning*. Birmingham, UK: Packt Publishing, 2015. ISBN: 1783555130.

[46] LM Rebull, S Guieu, JR Stauffer, LA Hillenbrand, A Noriega-Crespo, KR Stapelfeldt, SJ Carey, JM Carpenter, DM Cole, DL Padgett, et al. "The North American and Pelican Nebulae. II. MIPS Observations and Analysis". In: *The Astrophysical Journal Supplement Series* 193.2 (2011), p. 25.

[47] Joseph W Richards, Peter E Freeman, Ann B Lee, and Chad M Schafer. "Exploiting low-dimensional structure in astronomical spectra". In: *The Astrophysical Journal* 691.1 (2009), p. 32.

[48] Joseph W Richards, Darren Homrighausen, Peter E Freeman, Chad M Schafer, and Dovi Poznanski. "Semi-supervised learning for photometric supernova classification". In: *Monthly Notices of the Royal Astronomical Society* 419.2 (2011), pp. 1121–1135.

[49] Paul Robertson, Michael Endl, William D Cochran, and Sarah E Dodson-Robinson. "Hα activity of old M dwarfs: Stellar cycles and mean activity levels for 93 low-mass stars in the solar neighborhood". In: *The Astrophysical Journal* 764.1 (2013), p. 3.

[50] Arthur L. Samuel. "Some studies in machine learning using the game of Checkers". In: *IBM JOURNAL OF RESEARCH AND DEVELOPMENT* (1959), pp. 71–105.

[51] S Scaringi, C Knigge, JE Drew, M Monguió, E Breedt, M Fratta, B Gänsicke, TJ Maccarone, AF Pala, and C Schill. "The Gaia/IPHAS and Gaia/KIS value-added catalogues". In: *Monthly Notices of the Royal Astronomical Society* 481.3 (2018), pp. 3357–3369.

[52] Kevin Schawinski, Ce Zhang, Hantian Zhang, Lucas Fowler, and Gokula Krishnan Santhanam. "Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit". In: *Monthly Notices of the Royal Astronomical Society: Letters* 467.1 (2017), pp. L110–L114.

[53] *Scikit-Learn DBSCAN*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.dbscan.html.

[54] *Scikit-Learn python package*. URL: http://scikit-learn.org/.

[55] *Scipy python package*. URL: https://www.scipy.org/.

[56] Claude Elwood Shannon. "Communication in the presence of noise". In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21.

[57] VS Shevchenko, MA Ibragimov, and TL Chenysheva. "SFR 2-CYGNI-a Star-Forming Region Associated with the Extremely Young Cluster NGC6910 and the Herbig Be-Stars BD+ 40DEG4124 and BD+ 41DEG3731". In: *Soviet Astronomy* 35 (1991), p. 229.

[58] *Simbad - Guide*. URL: http://simbad.u-strasbg.fr/guide/simbad.htx.

[59] *Simbad Object Types*. URL: http://simbad.u-strasbg.fr/simbad/sim-display?data=otypes.

[60] Augustin Skopal. "How to understand the light curves of symbiotic stars". In: *arXiv preprint arXiv:0805.1222* (2008).

[61] Jorick S Vink, Janet E Drew, Danny Steeghs, Nick J Wright, Eduardo L Martin, Boris T Gänsicke, Robert Greimel, and Jeremy Drake. "IPHAS discoveries of young stars towards Cyg OB2 and its southern periphery". In: *Monthly Notices of the Royal Astronomical Society* 387.1 (2008), pp. 308–318.

[62] *Vizier*. URL: https://vizier.u-strasbg.fr/viz-bin/VizieR.

[63] Kyle W Willett, Chris J Lintott, Steven P Bamford, Karen L Masters, Brooke D Simmons, Kevin RV Casteels, Edward M Edmondson, Lucy F Fortson, Sugata Kaviraj, William C Keel, et al. "Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey". In: *Monthly Notices of the Royal Astronomical Society* 435.4 (2013), pp. 2835–2860.

[64] RE Williams. "Emission lines from the accretion disks of cataclysmic variables". In: *The Astrophysical Journal* 235 (1980), pp. 939–944.

[65] Andrew R Witham, C Knigge, BT Gänsicke, A Aungwerojwit, RLM Corradi, JE Drew, R Greimel, PJ Groot, L Morales-Rueda, ER Rodriguez-Flores, et al. "The properties of cataclysmic variables in photometric Hα surveys". In: *Monthly Notices of the Royal Astronomical Society* 369.2 (2006), pp. 581–597.

[66] AR Witham, C Knigge, JE Drew, R Greimel, D Steeghs, BT Gänsicke, PJ Groot, and A Mampaso. "The IPHAS catalogue of Hα emission-line sources in the northern Galactic plane". In: *Monthly Notices of the Royal Astronomical Society* 384.4 (2008), pp. 1277–1288.

[67] Olga V Zakhozhay, Anatoly S Miroshnichenko, Kenesken S Kuratov, Vladimir A Zakhozhay, Serik A Khokhlov, Sergey V Zharikov, and Nadine Manset. "IRAS 22150+ 6109–a young B-type star with a large disc". In: *Monthly Notices of the Royal Astronomical Society* 477.1 (2018), pp. 977–982.

[68] Michael Zevin, Scott Coughlin, Sara Bahaadini, Emre Besler, Neda Rohani, Sarah Allen, Miriam Cabero, Kevin Crowston, Aggelos K Katsaggelos, Shane L Larson, et al. "Gravity Spy: integrating advanced LIGO detector characterization, machine learning, and citizen science". In: *Classical and quantum gravity* 34.6 (2017), p. 064003.

[69] Gang Zhao, Yong-Heng Zhao, Yao-Quan Chu, Yi-Peng Jing, and Li-Cai Deng. "LAMOST spectral survey—An overview". In: *Research in Astronomy and Astrophysics* 12.7 (2012), p. 723.